中国科学: 物理学 力学 天文学 2011 年 第 41 卷 第 11 期: 1342 ~ 1350
SCIENTIA SINICA Phys, Mech & Astron www.scichina.com phys.scichina.com
《中国科学》杂志社
SCIENCE CHINA PRESS

论 文

应用数据挖掘技术的短期太阳耀斑预报模型

李蓉[①*], 崔延美[②]

① 北京物资学院, 北京 101149
② 中国科学院空间科学与应用研究中心, 北京 100190
*E-mail: lirong@bao.ac.cn

收稿日期: 2011-06-09; 接受日期: 2011-09-05; 网络出版日期: 2011-10-19
国家自然科学基金(编号: 10973020)、北京市属高等学校人才强教计划资助项目(编号: PHR200906210)、北京市教育委员会科研基地建设项目(编号: WYJD200902)和北京市教育委员会科技计划项目(编号: KM200810037001)资助

摘要 为了进一步探讨太阳耀斑与太阳黑子参量的关系, 本文采集了大规模的活动区黑子数据, 统计其与耀斑发生的产率关系, 应用得到的拟和公式对原始数据计算得到规范化后的数据集. 在此基础上使用数据挖掘技术对黑子耀斑数据建立决策树模型和建立分类规则, 具体描述了黑子数据和太阳耀斑之间的相关性. 最后应用这两种技术对活动区未来 48 h 是否爆发耀斑给出了预报, 预报结果具有较高的准确率和较低的虚报率.

关键词 太阳耀斑, 黑子数据, 预报因子, 决策树, 分类规则

PACS: 96.60.Qe, 96.60.Q-, 92.70.Qr

1 引言

太阳耀斑是最剧烈的太阳活动现象之一, 耀斑事件引起的 X 光辐射增强将破坏地球电离层的正常状态, 耀斑的高能粒子流将造成地球轨道附近高能粒子污染并干扰地球磁层, 向下传播导致地球低层大气热力学状态的变化. 通过这些扰动, 太阳耀斑对人类的航天活动、无线电通讯, 以及天气和水文领域产生影响. 因而预报太阳耀斑的发生又有实际应用的价值[1].

在太阳耀斑预报技术研究中, 建立耀斑和太阳活动区特征参量的关系十分重要. 研究表明黑子和耀斑发生有着紧密的联系, 已有的一些太阳耀斑预报模型采用黑子参量作为预报因子. NOAA 空间环境实验室和 Colorado 大学联合开发了一个专家系统 Theo 来预报太阳耀斑[2]. Theo 选择黑子群的 McIntosh 分类作为主要的知识基, 对耀斑和黑子的联系进行了编码, 建立了一个强大的耀斑及分类的数据库和一个通过知识推理机的附加预测规则. 美国大熊湖天文台的 Gallagher 等人[3]根据第 22 太阳活动周开始以来大约 8 年的 NOAA 资料, 发展了一种太阳耀斑的预报方法, 根据 McIntosh 分类的三个参量(Z, p, c)进行预报, 对每一个活动区产生 C 级、M 级、或 X 级的 X 射线耀斑的可能性进行预报. 悉尼大学的 Wheatland 提出了用贝叶斯方法进行耀斑预报[4]. 这个方法使用一个活动区的耀斑发生记录和耀斑统计的现象规则对大耀斑在子序列时间段的发生情况产生一个最初的预测. 这个初始的预测再通过一个现存的耀斑预报方法来评价. Qahwaji 和 Colak[5]提出了应用机器学习方法建立了短期太阳耀斑预报模型,

引用格式: 李蓉, 崔延美. 应用数据挖掘技术的短期太阳耀斑预报模型. 中国科学: 物理学 力学 天文学, 2011, 41: 1342–1350
Li R, Cui Y M. Short-term solar flare forecasting model using data mining technique (in Chinese). Sci Sin Phys Mech Astron, 2011, 41: 1342–1350, doi: 10.1360/132011-710

模型采用黑子群的 McIntosh 分类的三个参数 Z, p, c 和黑子数作为输入参量并比较了几种机器学习方法的耀斑预报准确率.

此外, 国家天文台太阳活动预报中心在太阳耀斑预报中做了系列研究工作. 早期包括张桂清老师对太阳黑子数据与耀斑关系进行分区间统计, 应用多元判别的方法建立了太阳耀斑预报模型[6], 该模型已应用于太阳活动预报中心的每日太阳耀斑预报工作. Zhu 等人[7]对采用这种方法的预报结果给出了统计, 并和美国 WWA 的预报结果进行了比对, 结果显示该方法的预报准确率高于 WWA 的预报结果. 基于这些黑子统计数据, 文献[8]提出了应用支持向量机和近邻相结合的方法建立耀斑预报模型, 并比较两种预报方法的预报准确率. 研究工作除了采用活动区黑子参量作为预报因子, 还将磁场参量引入到太阳耀斑预报中. 文献[9]对纵向磁场最大水平梯度, 强梯度中心线长度, 孤立奇点数目三个物理参量进行了计算, 统计磁场参量与耀斑的产率拟合函数, 为预报耀斑奠定了基础. 基于这些磁场参量数据, Wang 等人提出了使用神经网络建立耀斑预报模型, 给出了较好的预报结果[10]. 文献[11]将时间序列信息引入到磁场数据, 采用学习矢量量化和聚类方法建立耀斑预报模型.

本文进一步探讨论黑子参量和耀斑之间的联系. 为了使统计分析结果更具客观性, 我们的工作建立在大规模的数据集上, 数据范围覆盖了第 23 太阳活动周. 采集连续 12 y 的太阳活动区黑子和耀斑数据, 计算耀斑和黑子数据之间的产率, 建立它们之间的拟合关系. 在此基础上, 应用数据挖掘技术对数据进行分析. 分别用决策树技术和建立分类规则挖掘黑子和耀斑发生之间的关联模式, 并对未来 48 h 是否发生耀斑给出了预报结果.

2 数据介绍

太阳 X 射线耀斑数据来自 GOES 的观测资料, 可由网址 ftp://ftp.ngdc.noaa.gov/STP/SOLAR_DATA/下载. 耀斑通常根据峰值流量 1~8 Aº分为 B, C, M, X 四大级别, 其中 C-, M-, X-级对地球产生影响.

考虑到所选活动区产生至少一个 C1.0 的耀斑, 我们以 C1.0 作为基本单位, 对一定时间内发生的所有耀斑定义一个耀斑总级别 $I_{\rm tot}$, 即

$$I_{tot} = 0.1\times\sum B + 1.0\sum C + 10\times\sum M + 100\times\sum X. \quad (1)$$

如某一活动区在 48 小时内爆发了五个耀斑, 其级别分别是 B9.6, C1.2, C2.3, M4.1, X1.2, 那么耀斑总级别为 $I_{\rm tot}$ = 0.1×9.6+(1.2+2.3)+10×4.1+100×1.2= 165.46. 一个可操作的耀斑预报模型通常关注在未来一定时间内大于某一级别的耀斑发生情况. 本文关注的耀斑为总级别 $I_{\rm tot}\geqslant 10$, 即相当于一个 M1.0 的耀斑[9].

太阳黑子数据来源于 NOAA 的每日活动区总结报告 1). 报告中记录了每日黑子活动区的太阳观测, 包括编号, 位置, 面积和分类情况. 黑子主要存在两个分类系统, 分别为 Wilson 山磁分类和 McIntosh 分类. Wilson 山磁分类是 1919 年美国 Wilson 山天文台提出黑子群按磁场极性分类方法, 它以双极黑子为一个基本类型, 其他类型都看作双极黑子群的变形. McIntosh 分类是修改的 Zurich 分类, 根据黑子群演化过程中所表现出的形态进行分类. McIntosh 分类通常的形式是 Zpc, Z 是修改的 Zurich 分类, p 是黑子类型, c 是黑子组分别间隔的紧凑度[1]. 在我们的数据分析中选取三个黑子参量分别为黑子面积、磁分类和 McIntosh 分类. 考虑到 10.7 cm 射电流量和太阳活动的变化密切相关, 我们还选择了太阳射电流量 2).

3 黑子参量与耀斑产率的关系

上述耀斑与黑子数据组成了原始的数据集, 活动区的太阳黑子数据作为数据集的输入, 活动区未来 48 h 的耀斑发生情况作为输出. 输入中黑子的磁分类、McIntosh 分类属于取值离散的变量采用一种统计方法, 而黑子群面积和 10 cm 射电流量作为连续型变量应用另一种统计方法, 下面分别给以描述.

3.1 黑子的磁分类、McIntosh 分类与耀斑产率

我们对黑子的磁分类与 McIntosh 分类的每一类型计算它与耀斑的产率, 统计方法为取数据集这一类型中耀斑爆发的个数与该类型总样本数的比值.

1) http://www.swpc.noaa.gov/ftpmenu/forecasts/SRS.html
2) http: ftp://ftp.ngdc.noaa.gov/STP/SOLAR_DATA/

例如对于McIntosh分类中的Ekc, 属于这个类型的样本总数为230, 其中爆发M级耀斑的个数为80个, 那么Ekc类型的耀斑产率为80/230=0.35. 两个参量的耀斑产率统计结果见表1和2, 每个分类的活动区总样本数与发生耀斑的活动去样本数见图1, 统计的发生耀斑产率见图2. 由图2可以看出黑子磁分类的γ分类和McIntosh分类中的'Fhc'的耀斑产率最高.

3.2 黑子群面积、10 cm射电流量与耀斑产率

黑子群面积和10 cm射电流量的统计产率公式[9]如下所示

$$P(X) = S_a(X) / S_t(X), \tag{2}$$

(2)式中当预报因子的取值在[X, ∞]时$S_a(X)$是发生耀斑的样本数, $S_t(X)$是所有样本的个数. 例如当黑子面积取值为900时, 有217个样本的面积大于900, 其中爆发M级耀斑的个数为134个, 那么面积为900时的耀斑产率是134/217=0.62. 求出黑子群面积和10 cm射电流量对耀斑的产率后用函数拟和它们和耀斑产率的关系, 拟和函数为高斯函数:

$$Y = A_1 + \frac{A_2}{\sqrt{2\pi}W}\exp-\frac{(X - X_0)^2}{2W^2}, \tag{3}$$

拟和函数的参数见表3, 分段区间的总样本数和发生耀斑样本数及拟和曲线见图3.

4 数据挖掘技术

决策树和分类规则是两个主要的两个数据挖掘工具, 除了具有一般的机器学习方法的分类功能, 还

表1 磁分类与太阳耀斑的产率

Table 1 Magnetic class and solar flare productivity

磁分类	β	α	$\beta\delta$	$\beta\gamma$	$\beta\gamma\delta$	$\gamma\delta$	γ
耀斑产率	0.041	0.009	0.284	0.196	0.499	0.7	0.75

表2 McIntosh分类与太阳耀斑的产率

Table 2 McIntosh class and solar flare productivity

McIntosh 分类	Bxo	Axx	Cro	Dao	Hsx	Cso	Dso	Cao	Hrx	Cri	Dai	Eac	Ehi	Eho	Esi
太阳耀斑产率	0.009	0.006	0.013	0.062	0.007	0.021	0.036	0.036	0.003	0.5	0.131	0.151	0.259	0.138	0.079
McIntosh 分类	Eki	Eso	Dsi	Ekc	Dro	Eai	Eao	Hax	Dri	Fsi	Dko	Fki	Eko	Fko	Hkx
太阳耀斑产率	0.224	0.055	0.089	0.348	0.019	0.175	0.091	0.024	0	0.111	0.172	0.377	0.182	0.217	0.021
McIntosh 分类	Hhx	Fkc	Fsc	Fhi	Cho	Esc	Cko	Fai	Fho	Dho	Dac	Fhc	Fao	Dki	Cki
太阳耀斑产率	0.033	0.587	0.333	0.4	0.067	0.167	0.112	0.225	0.176	0.06	0.283	0.6	0.194	0.239	0.286
McIntosh 分类	Bxi	Cai	Dsc	Fso	Fac	Chi	Dkc	Ehc	Ero	Dhi	Csi	Fro	Dhc	Ckc	Hxx
太阳耀斑产率	0	0.087	0	0.082	0.323	0	0.367	0	0	0.111	0.125	0	0.333	0	0

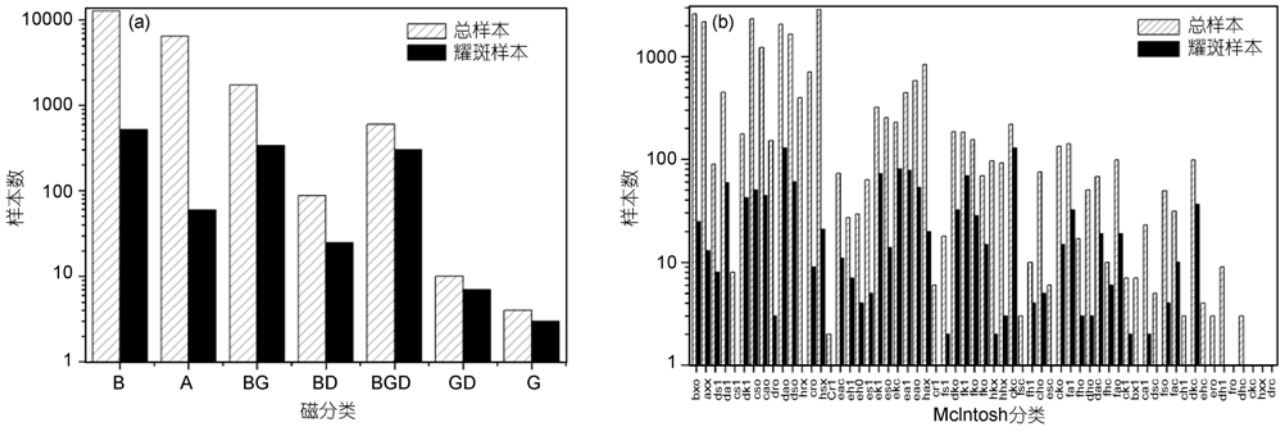

图1 磁分类(a)与McIntosh(b)的总样本与耀斑样本柱状分布图

Figure 1 Total and flare samples distribution of magnetic (a) and McIntosh class (b).

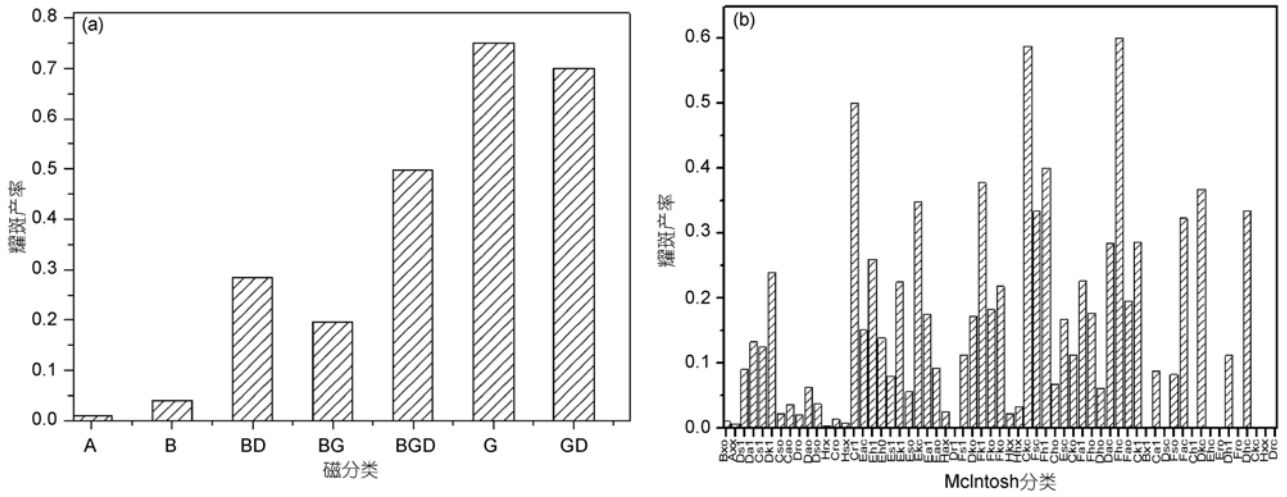

图 2　磁分类(a)与 McIntosh 分类(b)与耀斑产率

Figure 2　Flare productivity of magnetic class (a) and McIntosh class (b).

表 3　黑子面积和太阳射电流量的拟和参数值

Table 3　Fitting parameters value of sunspot and solar flux

预报因子	*A*1	*A*2	*X*0	*W*
黑子面积	–0.3933	3672.86	1636.15	2345.97
太阳射电流量	0.205	129.952	248.61	177.02

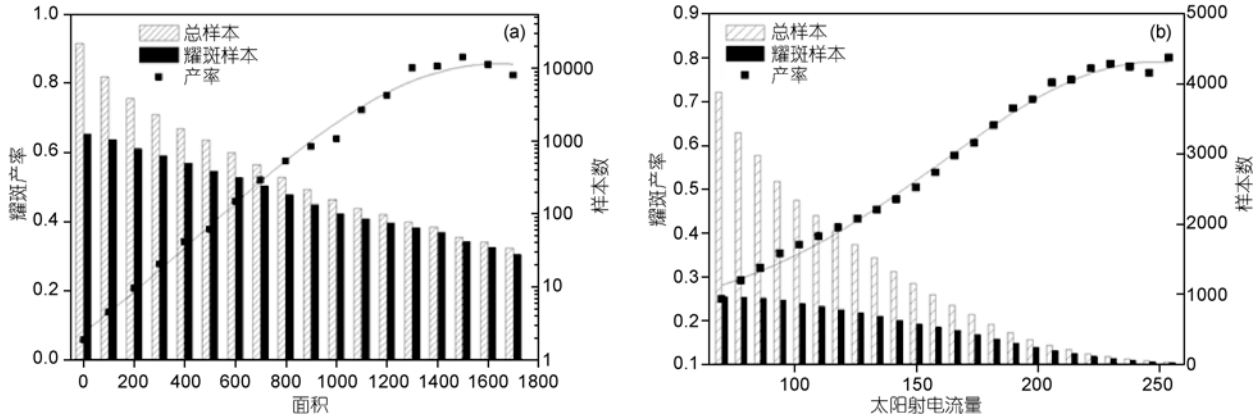

图 3　黑子面积(a)和太阳射电流量(b)的样本数和耀斑产率

Figure 3　Flare productivity of sunspot (a) and solar flux (b).

能够挖掘数据集中隐含的规则. 其中决策树是一种较常用的算法, 主要用于模式分类. 而分类规则主要用于数据挖掘领域, 得到的规则以“IF-THEN”规则的形式给出, 较好的表示了数据集输入输出之间的相关性.

4.1　决策树

决策树是以实例为基础的归纳学习算法. 构造决策树的方法采用自顶向下递归地分治方式构造, 按度量标准选择分裂属性, 在决策树的内部节点进行属性值的比较, 并根据不同的属性值从该节点向下分治, 叶节点是要学习划分的类. 从根到叶节点的一条路径就对应着一条分类规则. 当决策树模型建立后, 对一个未知实例进行分类时, 是从决策树的根节点出发, 根据在各个连续节点上对未知实例的属性值测试的结果, 自上而下地从树上寻找出一条路径, 当实例到达叶子时, 实例的分类就是叶子节点所

标注的类[12].

在构建决策树的过程中, 最重要的是定义属性重要度的评价指标, 按此指标选择分裂属性. 其中最常用的就是使用信息增益作为属性选择度量. D 为标注类的训练集, 假定有 m 个不同的类 C_i (i= 1, 2, …, m), C_{id} 是 D 中 C_i 类样本的集合, $|D|$和$|C_{id}|$分别是 D 中和 C_{id} 中样本的个数. D 中样本分类所需的期望信息由下式给出

$$\mathrm{Info}(D) = -\sum_{i=1}^{m} p_i \log_2(p_i), \tag{4}$$

式中 P_i 是 D 中任意样本属于类 C_i 的概率, 用$|C_{id}|/|D|$估计, Info(D)又称 D 的熵. 假设按属性 A 划分 D 中的样本, 属性 A 根据训练数据的观测具有 v 个不同的值 $\{a_1, a_2, \cdots, a_v\}$. 可以用属性 A 将 D 划分为 v 个子集 $\{D_1, D_2, \cdots, D_v\}$, 其中 D_j 包含 D 中的样本, 它们在 A 上具有值 a_j, 这些划分对应从节点 N 生长出的来得分支. 定义下式度量:

$$\mathrm{Info}_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times \mathrm{Info}(D_j). \tag{5}$$

$\mathrm{Info}_A(D)$是基于按 A 划分对 D 中样本分类所需要的期望信息. 所需要的期望信息越小, 划分的纯度越高. 信息增益定义为原来的信息需求和新的需求的差, 即

$$\mathrm{Gain}(A) = \mathrm{Info}(D) - \mathrm{Info}_A(D). \tag{6}$$

选择具有最高信息增益 Gain(A)的属性 A 作为节点 N 的分裂属性, 使得完成样本分类所需要的信息最小.

1986 年 Quinlan 提出了著名的 ID3 算法. 在 ID3 算法的基础上, 1993 年 Quinlan 又提出了 C4.5 算法[13]. ID3 算法的核心是: 在决策树各级结点上选择属性时, 用信息增益(Information Gain)作为属性的选择标准, 以使得在每一非叶子结点进行测试时, 能获得关于被测试记录最大的类别信息. 其具体方法是: 检测所有的属性, 选择信息增益最大的属性产生决策树结点, 由该属性的不同取值建立分支, 再对各分支的子集递归调用该方法建立决策树结点的分支, 直到所有子集仅包含同一类别的数据为止. 最后得到一棵决策树, 它可以用来对新的样本进行分类.

C4.5 算法继承了 ID3 算法的优点, 并在以下几方面对 ID3 算法进行了改进:

(1) 用信息增益率来选择属性, 克服了用信息增益选择属性时偏向选择取值多的属性的不足; 增益率 GainRatio 定义为

$$\mathrm{GainRatio}(A) = \frac{\mathrm{Gain}(A)}{\mathrm{SplitInfo}(A)}, \tag{7}$$

式中

$$\mathrm{SplitInfoA}(D) = -\sum_{j=1}^{v} \frac{|D_j|}{|D|} \times \log_2\left(\frac{|D_j|}{|D|}\right),$$

表示通过将训练集 D 划分成对应于属性 A 测试的 v 个输出的 v 个划分产生的信息.

(2) 在树构造过程中进行剪枝;

(3) 能够完成对连续属性的离散化处理;

(4) 能够对不完整数据进行处理.

C4.5 算法与其它分类算法如统计方法、神经网络等比较起来有如下优点: 产生的分类规则易于理解, 准确率较高.

4.2 分类规则

规则是表示信息或少量知识的一个有效办法, 分类规则是从实例中学习一组 IF-THEN 规则, 规则之间是析取(逻辑或)关系. 一个 IF-THEN 规则是一个如下表示的表达式:

IF 条件 THEN 结论 规则的"IF"部分称作规则前件或前提. "THEN"部分是规则的结论. 在规则前件, 条件由一个或多个用逻辑连词 AND 连接的属性测试(如)组成. 规则的结论是一个类预测.

规则可以从决策树中提取或使用顺序覆盖法直接从训练数据中提取. 本文使用顺序覆盖算法, 直接从训练数据提取 IF-THEN 规则. 顺序覆盖法是最广泛使用的挖掘分类规则吸取集的方法, 可以搜索数据中频繁出现的属性-值对. 这些属性-值对形成关联规则, 可以分析数据集并用于分类. 有许多顺序覆盖算法, 流行的算法包括 AQ, CN2 和最近提出的 RIPPER. 算法一般策略为一次学习一个规则, 每当学习完一个规则, 就删除该规则覆盖的实例, 并对剩下的实例重复该过程[13].

算法: 顺序覆盖. 学习一组 IF-THEN 分类规则.

输入: D, 类标记的实例的数据集合; Att_vals, 所有属性与它们的可能值的集合.

输出: IF-THEN 规则的集合.

方法:

(1) Rule_set={}

(2) For 每个类 c do

(3) Repeat
(4) Rule=Learn_one _Rule(D, Att-val,c);
(5)从 *D* 中删除 Rule 覆盖的实例;
(6) Until 终止条件满足;
(7) Rule_set=Rule_set+Rule;
(8) Endfor
(9) 返回 Rule_Set;

基本顺序覆盖算法中一次为一个类学习规则. 给定当前的数据集, Lear_one_Rule 是找出当前类的“最好”规则, 重复该过程直到满足终止条件, 如不再有训练实例或返回的规则的值低于用户指定的阈值. Lear_one_Rule 的一个通用算法如下[12]:

(1) 对于每个类别 *C*;
(2) 当实例集 *D* 中含有类别 *C* 的实例时;
(3) 创建规则 *R*, 规则左边为空条件, 预测类别为 *C*;
(4) 循环工作, 直到 *R* 完美(或没有属性可使用);
(5) 对每个在规则 *R* 中还未包含的属性 *A* 和属性值 *v*;
(6) 考虑添加条件 *A*=*v* 到规则 *R* 的左边;
(7) 选择 *A* 和 *v*, 使正确率最大化;
(8) 将 *A*=*v* 添加到规则 *R* 中;
(9) 将规则 *R* 所覆盖的实例从 *D* 中删除.

5 应用与结果

5.1 应用

数据集样本覆盖太阳第 23 活动周, 范围从 1996 年 1 月至 2008 年 12 月. 活动区的黑子参量和射电流量的原始数据经过第 3 节的产率转换得到规范化的数值, 结果作为样本集的输入, 被称为预报因子和实例的属性; 该活动区未来 48 h 是否爆发耀斑作为样本集的输出, 爆发耀斑作为正例样本标记 1, 未发生耀斑作为反例样本标记为 0. 表 4 显示数据集样本的 3 个例子, 每个例子的第一行表示预报因子和发生耀斑的原始数据, 第二行是规范后的数据, 用于建立决策树和分类规则.

数据集共有样本 21655 个样本, 其中发生耀斑的活动区样本数是 1252, 未发生耀斑样本数为 20404. 为了均衡样本集中正反例样本的比例, 我们采用下采样的方法, 即从反例样本中随机抽取和正例样本数量相同的样本. 这样得到数据集共 2504 个样本.

对于建立的数据集应用 WEKA(Waikato Environment for Knowledge Analysis)数据挖掘软件, 可由网址 http://www.cs.waikato.ac.nz/ml/weka/下载. 其中决策树算法采用 C4.5 算法, 在 WEKA 中称为 J48 算法, 分类规则采用 Ripper 算法. 两种算法中参数均采用默认的设置.

5.2 数据挖掘结果

数据实验采用十折交叉验证, 首先将数据集分成十份, 轮流将其中 9 份作为训练数据, 1 份作为测试数据. 每次试验都会得出相应的评价指标. 最后, 将 10 次评价指标的平均值作为对模型性能的估计. 建立的决策树模型和分类规则显示于图 4 和 5.

图 4 是一颗决策树模型, 数中的中间结点用椭圆型表示, 里面标注的是属性的名字, 每个结点有两个分支, 分别对应着对属性的测试结果, 路径上标的是测试的属性值. 根结点是‘McIn’表示 McIntosh 分类, 通过信息增益计算的结果作为首选的属性, 其余依次是磁分类‘Wils’, 面积‘Area’和射电流量‘Flux’. 叶子结点对应的类判别用巨型表示, ‘class 0’表示未发生耀斑样本类, ‘class 1’表示发生耀斑样本类.下面的数字表示此条路径的分类结果. 例如决策树中最左边的路径‘class 0 919/142’表示符合属性测试条件的 919 个反例样本中错分 142 个, 其概率为 84.5%.

表 4 数据集输入输出的样本实例

Table 4 Dataset input and output sample instances

No.	输入				输出	
	面积	McIntosh 分类	磁分类	射电流量	发生耀斑	耀斑级别
1	10	BXO	B	75.1	no	no flare
	0.04	0.08453	0.009	0.29044	0	0
2	250	DAI	BGD	81.6	Flare	X-class
	0.499	0.22811	0.131	0.30449	1	1
3	690	EKI	B	98.1	Flare	M-class
	0.04	0.50897	0.223	0.34259	1	1

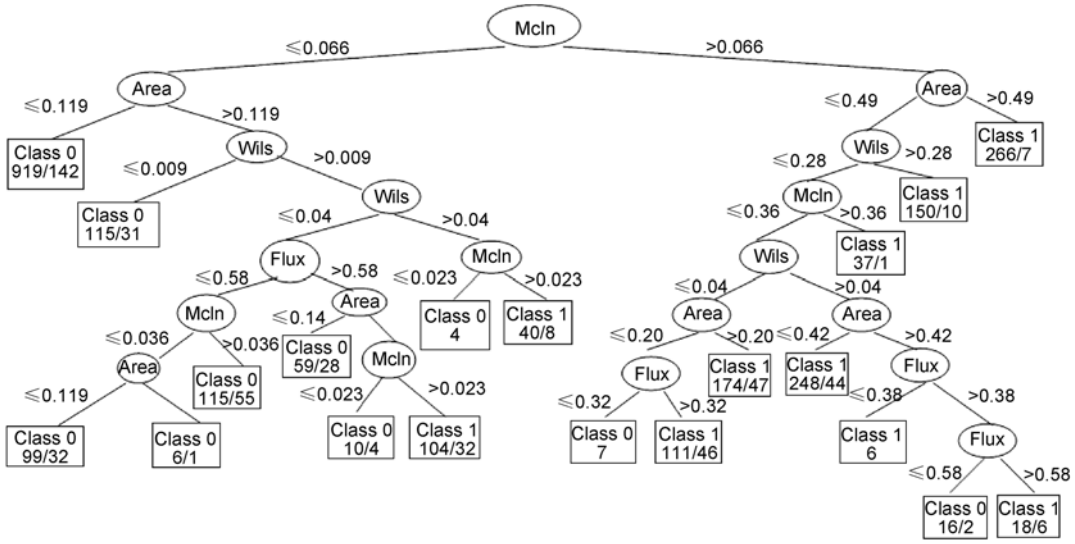

图 4　C4.5 决策树模型

Figure 4　Decision tree module.

①IF (McIn>= 0.066) and (Wils >= 0.195) Then class=1 (725.0/81.0)
②IF (Area >= 0.1366) and (McIn >= 0.111) Then class=1 (215.0/60.0)
③IF (Area >= 0.1307) and (Flux >= 0.59176) and (McIn >= 0.035) and (Flux >= 0.74005) Then class=1 (68.0/19.0)
④IF (McIn >= 0.036) and (Area >= 0.16048) Then class=1 (179.0/77.0)
⑤IF (Area >= 0.11898) and (Flux >= 0.6499) and (Wils >= 0.04) and (Flux <= 0.71172) Then class=1 (44.0/14.0)
⑥ => class=0 (1273.0/272.0)

图 5　Ripper 分类规则

Figure 5　Ripper class rule.

图 5 是用覆盖法建立的分类规则, 共建立了 6 条规则. 前 5 条表示对分类为 1 即发生耀斑的判定. 第一条规则的含义为“如果活动区的 McIntosh 分类 >=0.066 并且磁分类>=0.195 则发生耀斑”, 发生耀斑的概率 88.8%. 其余四条规则之间是析取关系, 第六条规则表示不满足上述五条分类规则的条件的剩余样本中,未发生耀斑的概率为 78.6%.

采用决策树和分类规则技术找出数据集输入和输出中存在的知识结构, 表现为决策树和分类规则的形式. 分类规则建立了黑子参量和耀斑发生直接的相关性联系, 表明当预报因子符合某类条件时发生或不发生耀斑. 这种关系可以给预报人员提供经验性的知识.

5.3　耀斑预报结果

决策树模型和分类规则除了提取预报因子和耀斑的关联模式, 还可以作为预报模型对每日活动区未来 48 h 是否发生耀斑给出预报. 在对预报结果进行评价时, 我们采用正确的肯定率(TP rate), 正确的否定率(TN rate), 虚报率(FP rate)和漏报率(FN rate)四个参量描述, 见表 5. 其中 TP 定义为对正例样本预报为正确的数目, TN 为对负类样本预报为正确的数目; FP 为对负类样本预报为正类的数目; FN 为对正类样本预报为负类的数目[14].

TP rate 为正确的正类样本数和真实的正类样本数的比:

$$\text{TP rate} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \tag{8}$$

TN rate 为正确的负类样本数与真实的负类样本

表 5　预报结果的表示

Table 5　Representation of prediction results

	预测的正类	预测的负类
真实的正类	TP	FN
真实的负类	FP	TN

数的比:

$$\text{TN rate} = \frac{\text{TN}}{\text{TN} + \text{FP}}. \quad (9)$$

FP rate 为虚报率, 定义为错误的否样本数和预报的正类样本数的比:

$$\text{FP rate} = \frac{\text{FP}}{\text{TP} + \text{FP}}. \quad (10)$$

FN rate 为漏报率, 定义为错误的肯定样本和真实的负类样本的比:

$$\text{FN rate} = \frac{\text{FN}}{\text{FP} + \text{TN}}. \quad (11)$$

模拟试验中我们将测试集的数据代入决策树模型和分类规则分别给出预报结果, 见表 6 和 7. 表 8 是由表 6 和表 7 计算出的预报准确率. 由表 8 的预报结果可以看出采用决策树和分类规则对测试集样本预报有较高的预报准确率. C4.5 决策树的 TPrate 和 TNrate 分别为 76.8%和 78.1%, Ripper 分类规则给出了略高的报准率(TPrate: 77.5%, TNrate: 79.2%). 而两个算法均有较低的虚报率, 分别为 22.2%和 21.1%.

表 6 C4.5 算法的预报结果

Table 6 Prediction result of C4.5

	预测的正类	预测的负类
真实的正类	961	291
真实的负类	274	978

表 7 Ripper 算法的预报结果

Table 7 Prediction result of C4.5

	预测的正类	预测的负类
真实的正类	970	282
真实的负类	260	992

表 8 C4.5 和 Ripper 算法的预报准确率显示

Table 8 Prediction accuracy of C4.5 and Ripper

	TP rate (%)	TN rate (%)	FP rate (%)	FN rate (%)
C4.5	76.8	78.1	22.2	23.2
Ripper	77.5	79.2	21.1	22.5

6 结论

本文主要研究太阳黑子数据和太阳耀斑之间的相关性统计分析. 本文首先采集了大规模的太阳活动区黑子耀斑数据,在此基础上完成了下列工作:

(ⅰ) 统计了太阳黑子参量和未来 48 h 发生耀斑之间的产率关系, 得到相关的拟和参数, 结果可作为黑子数据建立耀斑预报模型的数据预处理.

(ⅱ) 应用数据挖掘技术对计算完产率后的规范化数据集建立决策树模型和分类规则,描述了太阳耀斑与黑子数据间的相关性联系, 可为经验预报提供参考依据.

(ⅲ) 利用决策树和分类规则对测试集样本未来 48 h 发生耀斑的预报, 结果显示两种方法具有较高的预报准确率和较低的虚报率, 可以作为耀斑预报和其他太阳活动预报的有效方法.

前期的太阳耀斑预报研究工作已将太阳光球磁场参量进行了统计分析并应用于太阳耀斑预报[9~11], 下一步的工作是结合太阳活动区黑子数据和太阳光球磁场参量建立一个综合的太阳耀斑预报模型.

参考文献

1 林元章. 太阳物理学导论. 北京: 科学出版社, 2000

2 McIntosh P S. The classification of sunspot groups. Sol Phys, 1990, 125: 251–267

3 Gallagher P T, Moon Y, Wang H. Active-region monitoring and flare forecasting I. Data processing and first results. Sol Phys, 2002, 209: 171–183

4 Wheatland M S. A statistical solar flare forecast method. Space Weather, 2005, 7003–7011

5 Qahwaji R, Colak T. Automatic short-term solar flare prediction using machine learning and sunspot associations. Sol Phys, 2007: 241195

6 张桂清, 王家龙, 李德清. 短期 X 射线耀斑预报方案的详细介绍. 地球物理学进展, 1994, 9: 48–53

7 Zhu C L, Wang J L. Verication of short-term predictions of solar soft x-ray bursts for the maximum phase (2000-2001) of solar cycle 23. Chin J Astron Astrophys, 2003, 3: 563–568

8 Li R, He H, Cui Y M, et al. Support vector machine combined with K-Nearest Neighbors for solar flare forecast. Chin J Astron Astrophys, 2007, 7(3): 441–447

9 Cui Y M, Li R, Zhang L Y, et al. Correlation between solar flare productivity and photospheric magnetic field properties 1. Maximum

horizontal gradient, length of neutral line, number of singular points. Sol Phys, 2006, 237: 45–59
10 Wang H N, Cui Y M, Li R, et al. Solar flare forecasting model supported with artificial neural network techniques. Adv Space Res, 2007, 42: 1464–1468
11 Li R, Wang H N, Cui Y M, et al. Solar flare forecasting using learning vector quantity and unsupervised clustering techniques. Sci China Phys Mech Astron, 2011, 8: 1546–1552
12 Witten I H, Frank E. 数据挖掘-实用机器学习技术. 董琳, 邱泉, 译. 北京: 机械工业出版社, 2007
13 Han J W, Kamber M. 数据挖掘概念与技术. 范明, 孟小峰, 译. 北京: 机械工业出版社, 2006
14 Yu D R, Huang X, Wang H N, et al. Short-term solar flare prediction using a sequential supervised learning method. Sol Phys, 2009, 255: 91–105

Short-term solar flare forecasting model using data mining technique

LI Rong[1*] & CUI YanMei[2]

[1] *The Institution of Information, Beijing WuZi University, Beijing 101149;*
[2] *Center for Space Science and Applied Research, Chinese Academy of Sciences, Beijing 100190*

In order to study the correlation between the solar and sunspot parameters, a large scale of sunspots is collected, and the productivities with flare are calculated in this paper. After putting the original data into the fitting formula, the normalized data is obtained. Based on these data, decision tress and classification rule are constructed using data mining techniques, which describe concretely the correlation of sunspot and flare. Finally, these two modules give a forecasting about whether burst solar flare in region in future 48 hours. The results show a higher accuracy rate and a lower false alarm rate.

solar flare, sunspot data, predictor, decision tree, classification rule

PACS: 96.60.Qe, 96.60.Q-, 92.70.Qr

doi: 10.1360/132011-710