XIONG Senlin, LI Xinlu, FANG Shaofeng, ZOU Ziming. Research on solar proton event in the 23rd solar cycle using the machine learning methods (in Chinese). *Chin. J. Space Sci.*, 2021, **41**(3): 368-374. DOI:10.11728/cjss2021.03.368

基于机器学习方法对 23 太阳活动周质子事件的研究 *

熊森林　李昕璐　方少峰　邹自明

(中国科学院国家空间科学中心 空间大数据技术实验室　北京 100190)

摘　要　在耀斑伴随日冕物质抛射 (CME) 事件编目数据的基础上, 进行太阳质子事件 (SPE) 匹配, 构建研究数据集. 利用 Apriori 算法挖掘 SPE 与耀斑级别、耀斑发生日面位置以及 CME 角宽度和速度的关联关系. 结果表明: X 级耀斑、全晕 CME、高速 ($>1000\,\mathrm{km\cdot s^{-1}}$) CME 和日面西半球耀斑是最可能伴随质子事件的 4 种特征, 其诱发质子事件概率依次为 0.366, 0.355, 0.30, 0.155. 角宽度低于 120° 或速度低于 $400\,\mathrm{km\cdot s^{-1}}$ 的 CME 产生质子事件的概率为 0. 高速 CME 产生质子事件的概率是低速 ($400\sim1000\,\mathrm{km\cdot s^{-1}}$) CME 的 8.6 倍, X 级耀斑产生质子事件的概率是 M 级耀斑的 6.2 倍, 日面西部耀斑产生质子事件的概率是日面东部耀斑概率的 3.9 倍, 全晕 CME 产生太阳质子事件的概率是非全晕 (120°∼360°) CME 的 3.8 倍. 对太阳质子事件样本进行过采样处理, 利用随机森林等 5 种典型有监督学习算法, 构建了基于第 23 太阳活动周耀斑和 CME 特征的质子事件预测模型. 结果表明, 该预报模型的质子事件预测准确率、精确率和召回率均控制在 91% 以上.

关键词　太阳质子事件, 耀斑, 日冕物质抛射, Apriori 算法, 机器学习

中图分类号　P 353

Research on Solar Proton Event in the 23rd Solar Cycle Using the Machine Learning Methods

XIONG Senlin　LI Xinlu　FANG Shaofeng　ZOU Ziming

(*Space Science Big Data Technology Laboratory, National Space Science Center, Chinese Academy of Sciences, Beijing* 100190)

Abstract　Solar Proton Event (SPE) can pose crucial risks to the spacecraft. It is meaningful to analyze and build the relationships between SPE and the associated Coronal Mass Ejection (CME) and solar flares. In this study, the SPE in the 23rd solar cycle is investigated by using machine learning methods. Datasets were constructed based on CME and the solar proton events lists from 1997 to 2006 from the CDA web database. Apriori algorithm are used to survey the correlations between SPEs and the characteristics of flares and CME. The results show that X class flares, full halo CME, high speed (greater than $1000\,\mathrm{km\cdot s^{-1}}$) CME, and western flares are the four characteristics that most likely to be associated with SPE. The corresponding probabilities are 0.366, 0.355, 0.30 and 0.155. The

* 空间科学大数据管理与应用服务平台建设项目 (Z181100002918002) 和 "十三五" 信息化专项 (XXH13505-04) 共同资助
2019-12-09 收到原稿, 2020-08-24 收到修定稿

E-mail: xsl@nssc.ac.cn

SPE probabilities at the condition of more than one (CME or flare) features occurring simultaneously were exhibited as well. Using the over sampled CME and flares features, five SPE prediction models are built through five different supervised machine learning algorithms, thus Logistic Regression, Support Vector Classification, K-nearest neighbor, Random Forest and Gradient Boosting Decision Tree. The models all present pretty good prediction accuracy (> 0.94), precision (> 0.96) and recall rate (> 0.91).

Key words Solar Proton Event (SPE), Solar flares, Coronal Mass Ejection (CME), Apriori, Machine learning

0 引言

太阳耀斑爆发活动产生的巨大能量可将太阳粒子加速到 MeV 甚至 GeV. 这些能量粒子在行星际空间传播过程中被日冕物质抛射 (CME) 激发的行星际激波进一步加速, 形成地球轨道可观测的太阳高能粒子 (SEP). 空间环境预报业务通常将地球同步轨道位置连续 15 min 观测到能量大于 10 MeV 的高能质子通量超过 10 pfu 定为太阳质子事件 (SPE)[1]. SPE 是空间环境灾害事件之一, 这种突发的长时间 (数小时至数天)、高通量 (最高可达 10^5 pfu, 高出背景水平 5 个数量级) 的高能粒子辐射事件对在轨航天器安全运行、空间站航天员活动有潜在威胁. 因此, 对太阳质子事件爆发规律的研究和预测具有重要价值.

研究表明 [2], 太阳质子事件可分为脉冲型 (impulsive) 和渐进型 (gradual) 两种. 脉冲型质子事件由脉冲型耀斑加速, 粒子通量上升快, 事件持续时间短, 事件观测点和爆发源区通常位于磁力线连接的最佳位置. 渐进型质子事件强度大, 粒子通量上升缓慢, 事件持续时间长, 与持续时间较长的耀斑和日面较宽 ($> 120°$) 的快速 CME 相伴随. 尽管对太阳质子事件的分类还存在争议, 但普遍认为强质子事件由耀斑和 CME 产生的激波共同加速 [3]. Dierckxsens 等 [4] 对 23 太阳活动周的统计发现, 强耀斑、日面西部对应的 (耀斑) 爆发源区、高速 CME 以及全晕 CME 较其他条件有着更高的太阳质子事件爆发概率. Park 等 [5] 统计发现, 1976−2006 年期间, X 级耀斑引发质子事件的概率是 M 级耀斑的 10 倍; 30°−90°W 耀斑引发质子事件的概率是 30°−90°E 耀斑的 3 倍; 长时间 (> 0.3 h) 耀斑引发质子事件的概率是短时间 (< 0.3 h) 耀斑的 2 倍 (M 级) 和 7 倍 (X 级). Park 等 [6] 研究 1976−2006 年的质子事件发现, CME 速度、角宽度和位置条件对 CME 是否伴随质子事件有重要影响, 超过 1/3 的速度大于 1500 km·s^{-1} 的全晕 CME 伴随质子事件发生, 速度低于 1000 km·s^{-1} 的非全晕 CME 伴随质子事件的概率低于 1%.

上述研究主要运用传统的分组统计方法, 分析不同 CME 和耀斑条件下质子事件的发生概率. 利用 Apriori 算法准确而直接地挖掘质子事件与耀斑、CME 的关联关系未见报道. Apriori 算法是经典的机器学习算法之一, 可快捷地挖掘不同事物之间的关联, 即多个事物伴随发生的频繁度和伴随规则. 对太阳质子事件的预测/预警本质上是建立太阳质子事件在特定耀斑和 CME 条件下的模式分类, 而利用 CME 和耀斑特征构建的简单规则通常很难实现对质子事件的 100% 正确分类. 模型有较高报准率的同时, 虚报率往往也很高 [7]. 本文采用 Apriori 算法分析太阳质子事件和太阳爆发活动的关联性, 并基于 23 太阳活动周数据, 利用典型有监督机器学习算法训练太阳质子事件预测模型, 评估机器学习算法对太阳质子事件预测的表现.

1 数据与 Apriori 算法

本文研究用到的数据包括太阳质子事件编目数据* 和 Yashir[8] 整编的伴随 CME 的耀斑编目数据**. 耀斑和 CME 编目数据内容包括耀斑发生日期和时间、耀斑级别、耀斑日面位置, 以及耀斑伴随的 CME 角宽度、速度和中心角度信息. 这里仅选用耀斑伴随 CME, 且 CME 为标准格式的数据, 在 1997−2006

*https://cdaw.gsfc.nasa.gov/CME_list/sepe/

**https://cdaw.gsfc.nasa.gov/pub/yashiro/flare_cme/fclist_pub.txt

年期间的有效事件共计 679 个. 利用 NOAA 整编的质子事件数据, 对耀斑和 CME 数据进行质子事件匹配. 伴随质子事件的耀斑和 CME 标记为 1, 视为正样本; 不伴随质子事件的标记为 0, 视为负样本. 进行质子事件关联和标记后, 679 个耀斑和 CME 样本中共包括 68 个正样本 (伴随质子事件) 和 611 个负样本 (不伴随质子事件). 至此, 每个样本都包含了 CME 角宽度、速度及中心角度特征信息, 耀斑级别 (M 和 X)、经度 (东经为负, 西经为正) 特征信息, 以及是否伴随质子事件的标记信息.

在样本数据基础上, 对耀斑和 CME 特征进行分组和离散化处理. 表 1 给出了 CME 和耀斑的分组标准. 根据 CME 速度 $v_{\rm CME}$ 将 CME 划分为 0~400 $\rm km{\cdot}s^{-1}$, 400~1000 $\rm km{\cdot}s^{-1}$ 和 $>$ 1000 $\rm km{\cdot}s^{-1}$ 三组; 根据 CME 角宽度 $W_{\rm CME}$ 将 CME 划分为 [0, 120°), [120°, 360°), 360° (全晕) 3 组; 根据耀斑位置 $P_{\rm flare}$ 将耀斑分为日面东经 E (−90°−0) 和日面西经 W (0−90°) 两组; 根据耀斑强度 $f_{\rm flare}$ 将耀斑分为 M 级和 X 级 2 组; 根据耀斑和 CME 伴随质子事件状态 $S_{\rm SPE}$, 样本被划分为 1 (伴随) 和 0 (不伴随) 两种. 图 1 给出了在上述分类标准下, 各特征对应总样本和正样本统计分布.

表 1 耀斑和 CME 特征分组标准

Table 1 Discretization criteria for flare and CME features

Features	Feature groups
$v_{\rm CME}/(\rm km{\cdot}s^{-1})$	[0, 400), [400, 1000), [1000, +∞)
$W_{\rm CME}/(°)$	[0, 120), [120, 360), 360
$P_{\rm flare}/(°)$	E, W
$f_{\rm flare}$	M, X
$S_{\rm SPE}$	0, 1

本文利用 Apriori 算法中的频繁集和关联规则, 研究质子事件和耀斑、CME 特征的关联. 频繁集是一个或多个特征同时出现的集合, 用支持度 S 衡量. 频繁集的支持度是该频繁集在总样本中出现的频次, 求解方法如下:

$$S_{xy} = p_{xy} = \frac{N_{xy}}{N_{\rm t}}. \tag{1}$$

其中, p_{xy} 表示特征 x 和特征 y 同时出现的概率, N_{xy} 表示特征 x 和 y 同时出现的次数, $N_{\rm t}$ 为总样本数.

关联规则是频繁集中特征之间相互依存的关系, 用置信度 C (条件概率) 表示, 即某个特征成立的条件下, 其他特征成立的概率. 求解方法如下:

$$C(x|y) = p(x|y) = \frac{p(xy)}{p(y)}. \tag{2}$$

式中 $p(x|y)$ 为条件概率.

2 太阳质子事件与 CME 和耀斑的关联

利用 Apriori 算法, 分别计算了全样本和仅伴随质子事件的正样本中频繁集和关联规则. 基于全样本的一项频繁集支持度列于表 2. 频繁集结果显示, 太阳质子事件是小概率事件, 90% 的耀斑和其伴随的 CME 没有引发质子事件; M 级耀斑 (87%) 发生频度远高于 X 级耀斑 (13%); 出现在日面西部和东部的耀斑数量大致相当 (分别为 52% 和 48%). 超过

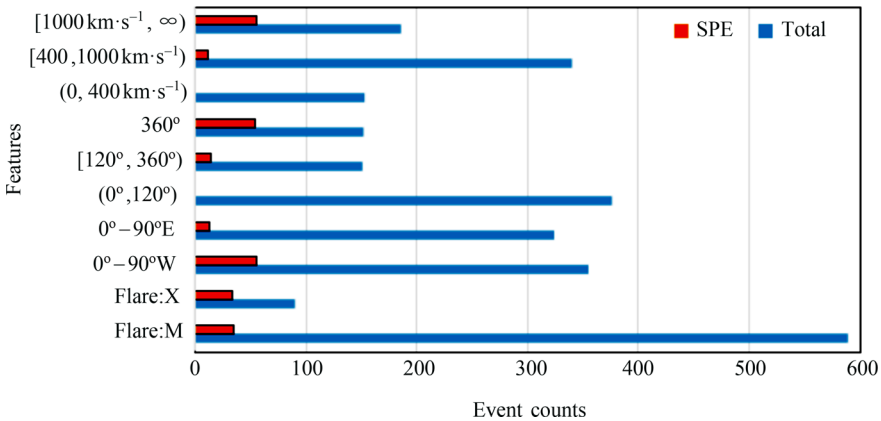


图 1 不同特征下总样本和正样本统计分布

Fig. 1 Distribution of total samples and SPE samples in different features

表 2 基于全样本的一项频繁集

Table 2 Frequent sets based on all sample dataset

频繁集	支持度 S
$\{S_{\rm SPE}$: 0$\}$	0.90
$\{f_{\rm flare}$: M$\}$	0.87
$\{W_{\rm CME}$: 0∼120°$\}$	0.55
$\{P_{\rm flare}$: W$\}$	0.52
$\{v_{\rm CME}$: 400∼1000 km·s$^{-1}\}$	0.50
$\{P_{\rm flare}$: E$\}$	0.48
$\{v_{\rm CME}$: $>$ 1000 km·s$^{-1}\}$	0.27
$\{v_{\rm CME}$: 0∼400 km·s$^{-1}\}$	0.23
$\{W_{\rm CME}$: 360°$\}$	0.22
$\{W_{\rm CME}$: 120°∼360°$\}$	0.22
$\{f_{\rm flare}$: X$\}$	0.13
$\{S_{\rm SPE}$: 1$\}$	0.10

半数 (55%) CME 角宽度小于 120°; 半数 (50%) CME 速度集中在 400∼1000 km·s^{-1}.

基于全样本的关联规则列于表 3. 当 CME 角宽度低于 120° 时, 不产生质子事件的概率为 100%; CME 角宽度介于 120°∼360° 时, 不产生质子事件的概率为 91%; 出现全晕 CME 时, 不产生质子事件的概率为 64%. CME 的速度低于 400 km·s^{-1} 时, 不产生质子事件的概率为 100%, 其关联的耀斑为 M 级耀斑的概率为 96%; CME 速度介于 400∼1000 km·s^{-1} 时, 不会发生质子事件的概率为 96%; 而当 CME 速度高于 1000 km·s^{-1} 时, 不产生质子事件的概率为 70%. 发生日面东部耀斑时, 不产生质子事件的概率为 96%, 耀斑位于日面西部时, 不产生质子事件的概率为 85%; 耀斑为 M 级时, 不产生质子事件的概率为 94%, 耀斑为 X 级时, 不产生质子事件的概率为 63%.

相对于所有样本频繁集和关联规则, 本文更关注伴随质子事件的 CME 和耀斑特征. 正样本中一项、两项和三项频繁集列于表 4.

正样本一项频繁集显示: 82% 的质子事件与速度大于 1000 km·s^{-1} 的高速 CME 相关, 剩余 18% 的质子事件与速度为 400∼1000 km·s^{-1} 的低速 CME 相关; 79% 的质子事件与全晕 CME 相关, 21% 的质子事件与角宽度介于 120°∼360° 的 CME 的相关; 81%

表 3 基于全样本的关联规则

Table 3 Association rules based on all sample dataset

关联规则	置信度 C
$\{S_{\rm SPE}$: 0$\}|\{W_{\rm CME}$: 0∼120°$\}$	1.0
$\{S_{\rm SPE}$: 0$\}|\{v_{\rm CME}$: 0∼400 km·s$^{-1}\}$	1.0
$\{S_{\rm SPE}$: 0$\}|\{v_{\rm CME}$: 400∼1000 km·s$^{-1}\}$	0.96
$\{S_{\rm SPE}$: 0$\}|\{P_{\rm flare}$: E$\}$	0.96
$\{S_{\rm SPE}$: 0$\}|\{f_{\rm flare}$: M$\}$	0.94
$\{S_{\rm SPE}$: 0$\}|\{W_{\rm CME}$: 120°∼360°$\}$	0.91
$\{S_{\rm SPE}$: 0$\}|\{P_{\rm flare}$: W$\}$	0.85
$\{S_{\rm SPE}$: 0$\}|\{v_{\rm CME}$: $>$ 1000 km·s$^{-1}\}$	0.70
$\{S_{\rm SPE}$: 0$\}|\{W_{\rm CME}$: 360°$\}$	0.64
$\{S_{\rm SPE}$: 0$\}|\{f_{\rm flare}$: X$\}$	0.63

的质子事件与日面西部耀斑相关, 仅 19% 的质子事件与日面东部耀斑相关; M 级质子耀斑和 X 级质子耀斑大致相当 (51% 和 49%).

正样本两项频繁集结果显示: 64% 的质子事件同时与全晕、高速 CME 相关; 63% 的质子事件同时与高速 CME 和日面西部耀斑相关; 60% 的质子事件同时与全晕 CME 和日面西部耀斑相关.

正样本三项频繁集结果显示: 接近一半 (46%) 的质子事件同时与全晕、高速 CME 和日面西部耀斑相关; 超过三分之一 (35%) 的质子事件同时与 X 级耀斑, 全晕、高速 CME 相关.

3 基于机器学习算法对质子事件的分类和验证

基于前面的分析可知, 耀斑位置 $P_{\rm flare}$、耀斑强度 $f_{\rm flare}$、CME 角宽度 $W_{\rm CME}$ 和 CME 速度 $v_{\rm CME}$ 是影响质子事件的重要参数. 这里利用耀斑和 CME 的 4 个特征, 采用机器学习领域中 5 个典型有监督分类模型, 建立并测试了质子事件分类模型效果.

3.1 数据处理

这里研究的数据集 (样本) 包含 68 个正样本 (伴随质子事件), 611 个负样本 (不伴随质子事件), 正负样本比例 1 : 9. 为不影响模型精度, 对太阳质子事件样本进行了过采样处理. 选用 SMOTE (Synthetic Minority Over-sampling Technique) 算法 [9] 对正样

表 4 基于伴随质子事件数据集的频繁集

Table 4 Frequent sets based on dataset associated SPE

	频繁集	支持度 S
一项频繁集	$\{v_{\rm CME}$: $>1000\,{\rm km\cdot s^{-1}}\}$	0.82
	$\{P_{\rm flare}$: W$\}$	0.81
	$\{W_{\rm CME}$: 360°$\}$	0.79
	$\{f_{\rm flare}$: M$\}$	0.51
	$\{f_{\rm flare}$: X$\}$	0.49
	$\{W_{\rm CME}$: 120°∼360°$\}$	0.21
	$\{P_{\rm flare}$: E$\}$	0.19
	$\{v_{\rm CME}$: 400∼1000 km·s$^{-1}\}$	0.18
二项频繁集	$\{W_{\rm CME}$: 360°, $v_{\rm CME}$: $>1000\,{\rm km\cdot s^{-1}}\}$	0.64
	$\{P_{\rm flare}$: W, $v_{\rm CME}$: $>1000\,{\rm km\cdot s^{-1}}\}$	0.63
	$\{P_{\rm flare}$: W, $W_{\rm CME}$: 360°$\}$	0.60
	$\{f_{\rm flare}$: X, $W_{\rm CME}$: 360°$\}$	0.43
	$\{f_{\rm flare}$: X, $v_{\rm CME}$: $>1000\,{\rm km\cdot s^{-1}}\}$	0.41
	$\{f_{\rm flare}$: M, $v_{\rm CME}$: $>1000\,{\rm km\cdot s^{-1}}\}$	0.41
	$\{P_{\rm flare}$: W, $f_{\rm flare}$: X$\}$	0.41
	$\{P_{\rm flare}$: W, $f_{\rm flare}$: M$\}$	0.40
	$\{P_{\rm flare}$: W, $W_{\rm CME}$: 360°$\}$	0.37
	...	
三项频繁集	$\{P_{\rm flare}$: W, $W_{\rm CME}$: 360°, $v_{\rm CME}$: $>1000\,{\rm km\cdot s^{-1}}\}$	0.46
	$\{f_{\rm flare}$: X, $W_{\rm CME}$: 360°, $v_{\rm CME}$: $>1000\,{\rm km\cdot s^{-1}}\}$	0.35
	$\{f_{\rm flare}$: X, $W_{\rm CME}$: 360°, $P_{\rm flare}$: W$\}$	0.34
	$\{f_{\rm flare}$: X, $v_{\rm CME}$: $>1000\,{\rm km\cdot s^{-1}}$, $P_{\rm flare}$: W$\}$	0.32
	$\{f_{\rm flare}$: M, $v_{\rm CME}$: $>1000\,{\rm km\cdot s^{-1}}$, $P_{\rm flare}$: W$\}$	0.31
	...	

本进行过采样处理, 以实现正负样本平衡.

具体步骤为: 从伴随质子事件的正样本中选取一个样本 N_i; 在所有正样本 T 中找到样本 N_i 的 k 个近邻, 并从这 k 个近邻中随机选择一个样本 M_{ik}; 利用随机生成的 0∼1 间的数作为 N_i 和 M_{ik} 距离权重, 即在这两个样本之间进行插值, 合成一个新的正样本; 针对样本 N_i 重复以上步骤 n 次, 合成 n 个正样本.

过采样处理后, 最终生成的数据集包含 611 个正样本, 负样本数量不变. 对新生成的样本, 采用 min-max 方法对数据进行标准化处理. 标准化后的数据按 7∶3 比例随机划分为训练集和验证集, 训练集用于建模, 验证集用于验证模.

3.2 质子事件分类模型的建立

3.2.1 逻辑回归模型

逻辑回归模型 (Logistic Regression, LG) 在狭义线性回归的基础上, 通过 Sigmoid 函数对 CME 和耀斑特征条件下的质子事件概率预测值进行转换, 得到 0∼1 之间的概率值; 之后通过阈值控制将概率值转换为质子事件和非质子事件两种类别, 从而实现对质子事件的分类.

3.2.2 支持向量机模型

支持向量机模型 (Support Vector Classification, SVC) 的核心思想是建立一个超平面, 使得两个不同类到该超平面最近点的距离最大化. 本文选用线性核函数和高斯核函数对输入的 CME 和耀斑数据进行特征空间映射, 通过设置核函数参数进行训练, 利用网格搜索和交叉验证找到精度最高的一组参数, 从而建立质子事件预测模型.

3.2.3 *K* 近邻模型

K 近邻模型 (K-Nearest Neighbor, KNN) 通过找出距离待分类样本最近的 k 个样本来实现分类. 影响 KNN 模型精度的三个因素为距离测算方法、类别决策方法的选取和 k 值设置. 本文采用欧几里得距

离进行距离测量, 利用投票法进行类别决策. 对于 k 值, 首先设置初始 k 值, 之后通过网格搜索和交叉验证得到训练精度最高的 k 参数.

3.2.4 随机森林模型

随机森林模型 (Random Forest, RF) 的基模型为决策树模型, 通过对原始数据集特征和样本随机抽样构建多棵决策树, 由投票法得到分类结果. 对基模型数量 (n_estimator) 和树最大深度 (max_depth) 两个参数进行了调整, 通过网格搜索和交叉验证得到最优参数, 建立质子事件预测模型.

3.2.5 梯度提升树模型

梯度提升树模型 (Gradient Boosting Decision Tree, GBDT) 的基模型也为决策树模型, GBDT 采用原始数据集的全部特征和全部样本建立基模型, 通过计算样本误差调整样本权重, 不断迭代生成多个基模型, 最终采用加权平均法得到分类结果. 本文对基模型数量 (n_estimator)、学习速率 (learning_rate) 和最大树深度 (max_depth) 进行调参, 通过交叉验证后得到质子事件预测的最佳参数.

4 模型分类结果

采用准确度 (accuracy)、精确度 (precision) 和召回率 (recall) 对以上 5 个模型进行精度验证. 准确度为质子事件和非质子事件均被正确分类占总样本的比例; 精确度为质子事件被正确分类占所有被分类为质子事件的比例; 召回率为真实质子事件中被正确分类的占比.

模型验证结果如图 2 所示. 由图 2 可以看出: 5 种模型验证的准确度、精确度和召回率均在 0.9 以上; 综合 3 项指标, 随机森林的精度最高.

5 讨论与结论

图 3 给出了不同 CME 和耀斑特征下太阳质子事件的概率 (特征条件的 SPE 数与特征条件总数的比值) 分布. 可以看出, 这里得到的质子事件概率密度与 Apriori 方法中的频繁集和关联规则结果一致. 质子事件概率值正向反映了特定条件下质子事件出现的可能性, 而全数据集的关联规则结果则直接给出了各特征条件下不发生质子事件的概率. 例如: 质子事件概率显示, M 级耀斑伴随质子事件的概率为

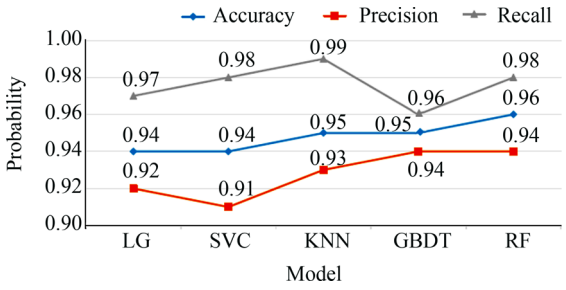


图 2 利用机器学习算法对太阳质子事件的分类结果

Fig. 2 Results of the SPE classification using different machine learning methods

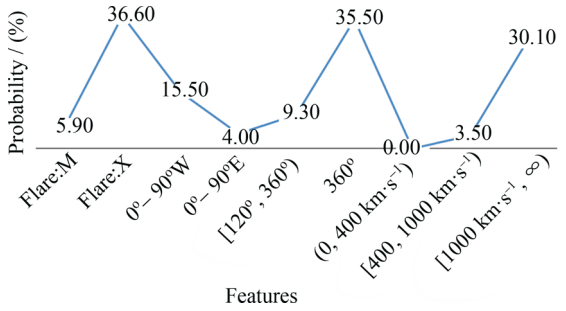


图 3 不同特征条件对应太阳质子事件的概率分布

Fig. 3 Distribution of SPE probabilities in different features

而全数据集的关联规则显示, M 级耀斑不产生质子事件的概率为 94% (保留 2 位有效数). 由此可知, 两种方法得出结果一致.

结合质子事件概率密度结果和 Apriori 方法频繁集和关联规则结果, 可以得出以下结论.

高速 ($> 1000\,\mathrm{km\cdot s^{-1}}$) CME 是产生质子事件的有利条件. 在所有样本中, 半数 (340 个) CME 的速度介于 $400\sim1000\,\mathrm{km\cdot s^{-1}}$, 这些 CME 仅伴随 12 个 (17.6%) 质子事件; 剩余 82.4% (56 个) 质子事件由高速 CME 产生. 而速度低于 $400\,\mathrm{km\cdot s^{-1}}$ 的 CME 不能产生质子事件. 高速 CME 产生质子事件的概率 (0.3) 是低速 ($400 < v < 1000\,\mathrm{km\cdot s^{-1}}$) CME 产生质子事件概率 (0.035) 的 8.6 倍.

强耀斑容易引发/伴随质子事件. X 级质子耀斑数量和 M 级质子耀斑数量相当 (33/35), 而 M 级耀斑数量是 X 级耀斑的 6.5 倍 (589/90). 换言之, X 级耀斑中, 质子耀斑占比 (0.366) 是 M 级质子耀斑占比 (0.059) 的 6.2 倍.

宽角度 CME 更利于质子事件的形成. 超过半数 (55%) 的 CME 角宽度低于 120°, 这些 CME 无一例外, 未能将行星际太阳风粒子加速到质子事件水平.

79.4% 质子事件伴随的 CME 为全晕 CME, 其余 20.6% 的质子事件伴随 CME 角宽度介于 120°~360°. 全晕 CME 产生质子事件的概率 (0.335) 是非全晕 CME 产生质子事件概率 (0.093) 的 3.8 倍.

发生于日面西部的耀斑更大概率为质子耀斑. 所有样本中, 发生于日面西部位置的耀斑占比 52.3%, 略高于日面东经耀斑. 但前者对应 80.8% (55 个) 的质子事件, 后者伴随的质子事件不到 20%. 西部耀斑为质子耀斑的概率 (0.155) 是东部耀斑为质子耀斑的概率 (0.04) 的 3.9 倍.

所有 10 组特征中, 强耀斑 (X 级以上)、全晕 CME、高速 CME 和日面西部耀斑是 4 组最容易产生质子事件的特征, 对应的质子事件概率分别为 0.366, 0.355, 0.3 和 0.155.

质子事件发生时: 4 组最容易产生质子事件的条件必定满足其一. 产生太阳质子事件时, 4 组容易产生太阳质子事件的条件成对出现的概率高, 4 组条件中至少满足两条的概率为 37%~64%, 至少满足三条的概率为 32%~46%.

本文分析结果与前人的研究结果基本一致. Dierckxsens 对 23 太阳活动周的分析结果显示, 强耀斑、出现在日面西部耀斑、高速 CME 和全晕 CME 较其他条件更有利于形成质子事件. Park 对 1976−2006 年数据的统计结果显示, X 级耀斑为质子耀斑的概率为 M 级耀斑为质子耀斑概率的 10 倍 (本文为 6.2 倍), 30°−90°W 耀斑引发质子事件的概率是 30°−90°E 耀斑引发质子事件概率的 3 倍 (本文西经耀斑伴随质子事件概率为东经耀斑伴随质子事件概率的 3.9 倍), 速度大于 1000 $\mathrm{km\cdot s^{-1}}$ 的全晕 CME 伴随质子事件的概率超过 1/3 (基于本文统计数据为 45%), 速度低于 1000 $\mathrm{km\cdot s^{-1}}$ 的非全晕 CME 伴随质子事件的概率低于 1% (基于本文的统计数据为 0.5%).

耀斑强度、耀斑位置、CME 速度和 CME 宽度是预警太阳质子事件的有效特征. 本文对质子事件样本进行过采样插值后, 通过机器学习算法实现特征数据的非线性映射处理, 再利用交叉验证找出最佳模型参数, 建立基于耀斑和 CME 特征的质子事件分类模型. 测试结果表明, 在对 23 太阳活动周数据训练中, 逻辑回归、支持向量机、K 近邻、梯度提升树和随机森林 5 种机器学习方法可有效地对质子事件进行分类, 模型的准确度、精度和召回率都在 91% 以上, 召回率可达 96% 以上.

致谢 国家科技资源共享服务平台——国家空间科学数据中心为本研究提供了计算服务和应用平台支持.

参考文献

[1] BALCH C C. SEC proton prediction model: verification and analysis [J]. *Radiat. Meas.*, 1999, **30**(3): 231-250

[2] DESAI M I, GIACALONE J. Large gradual solar energetic particle events [J]. *Living Rev. Sol. Phys.*, 2016, **13**(1): 1-132

[3] REAMES D V. The two sources of solar energetic particles [J]. *Space Sci. Rev.*, 2013, **175**(1): 53-92

[4] DIERCKXSENS M, TZIOTZIOU K, DALLA S, *et al.* Relationship between solar energetic particles and properties of flares and CMEs: statistical analysis of solar cycle 23 events [J]. *Sol. Phys.*, 2015, **290**(3): 841-874

[5] PARK J, MOON Y J, LEE D H, *et al.* Dependence of solar proton events on their associated activities: flare parameters [J]. *J. Geophys. Res.*, 2010, **115**(1): 39

[6] PARK J, MOON Y J, GOPALSWAMY N. Dependence of solar proton events on their associated activities: coronal mass ejection parameters [J]. *J. Geophys. Res.*, 2012, **117**(A8): 108-115

[7] XIONG Senlin, CUI Yanmei, LIU Siqing. Research on solar proton event warning with observation data of ACE satellite [J]. *Chin. J. Space Sci.*, 2013, **33**(4): 387-395 (熊森林, 崔延美, 刘四清. 利用 ACE 卫星数据对太阳质子事件预警方法的研究 [J]. 空间科学学报, 2013, **33**(4): 387-395)

[8] YASHIRO S, AKIYAMA S, GOPALSWAMY N, *et al.* Different power-law indices in the frequency distributions of flares with and without coronal mass ejections [J]. *Astrophys. J.*, 2006, **650**(2): 143-146

[9] CHAWLA N V, BOWYER K W, HALL L O, *et al.* Smote: synthetic minority over-sampling technique [J]. *J. Artif. Intell. Res.*, 2002, **16**(1): 321-357