结合主成分分析和支持向量机的太阳耀斑预报模型

袁飞^{①②}, 林佳本^{①*}, 邓元勇^①, 郭晶晶^①, 王刚^③, 曾真^①, 王丙祥^①

① 中国科学院太阳活动重点实验室, 北京 100012;

② 中国科学院大学, 北京 100049;

③ 北京交通大学, 北京 100044

* 联系人, E-mail: jiabenlin@bao.ac.cn

2015-12-28 收稿, 2016-01-20 修回, 2016-01-22 接受, 2016-05-17 网络版发表

国家自然科学基金(11373042, 11373044, 11427901)和中国科学院知识创新工程项目(KJCX2-EW-T07)资助

摘要 太阳耀斑是剧烈的太阳活动现象之一, 耀斑的预报对人类活动有着重要的实用价值. 为进一步提高太阳耀斑的预报准确率, 本文在综合考虑太阳黑子活动区参量、10.7 cm太阳射电流量等预报因子的前提下, 提出了结合主成分分析和支持向量机的太阳耀斑预报模型. 本模型的太阳黑子活动区参量包括黑子群面积、黑子群的McIntosh分类、活动区日面经度延伸、可见黑子数和黑子群的磁分类. 本文首先对上述参量进行了合适的属性编码并归一化建模所需数据集, 然后利用主成分分析方法提取出主要特征, 应用支持向量机方法建立了耀斑预报模型. 最后, 本文将该模型预报结果与其他预报模型的结果进行了对比, 结果验证了结合主成分分析和支持向量机的太阳耀斑预报模型是一种有效的预报模型.

关键词 太阳耀斑, 预报因子, 支持向量机, 主成分分析, 黑子群

太阳耀斑是剧烈的太阳活动现象之一^[1]. 剧烈的太阳活动对地球大气气压、大气电状态等具有明显的扰动作用^[2], 并且, 太阳耀斑爆发时释放出的大量高能粒子会影响地球磁场和其上空的电离层, 进而影响卫星导航、无线电通信等人类活动. 因此, 太阳耀斑的预报对人类活动有着重要的实用价值.

目前太阳耀斑的预报主要分为两大类: 基于物理知识的经验预报模型与基于观测数据的统计预报模型. 第一类预报模型是基于该领域内的专家提供耀斑预报知识而建立的物理预报模型, 该模型是在了解太阳耀斑发生过程与其变化机制的基础上而建立起来的预报模型. Cui等人^[3]计算并研究了纵向磁场最大水平梯度、磁中性线长度、孤立奇点数等磁场参量与太阳耀斑产率的关系, 该研究为建立基于磁场参量的太阳耀斑预报奠定了基础. Wang等人^[4]的

研究表明中性线附近的磁场梯度变化和磁剪切等参量与大耀斑的发生具有较大的相关性. 这些研究可为太阳耀斑预报模型的建立提供一定的理论指导, 然而, 由于目前太阳耀斑爆发的物理机制理解尚不彻底, 很难完全从物理的角度来预报耀斑的发生^[5]. 第二类预报模型是基于观测数据的处理与分析, 利用统计学和机器学习等方法建立的太阳耀斑预报模型. 研究表明, 太阳耀斑的发生和太阳黑子有着紧密联系^[6]. 20世纪80年代, NOAA空间环境实验室和Colorado大学联合开发了一个以黑子群的McIntosh分类作为主要知识基的专家系统来预报太阳耀斑^[7]. 之后, 美国大熊湖天文台根据McIntosh分类的3个参量(Z , p , c), 发展出了一种可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的太阳耀斑预报模型^[6]. 在国内, 中国科学院国家天文台通过

引用格式: 袁飞, 林佳本, 邓元勇, 等. 结合主成分分析和支持向量机的太阳耀斑预报模型. 科学通报, 2016, 61: 2316–2321

Yuan F, Lin J B, Deng Y Y, et al. Solar flare forecasting model of combining principal component analysis with support vector machine (in Chinese). Chin Sci Bull, 2016, 61: 2316–2321, doi: 10.1360/N972015-01342

统计太阳活动区黑子数据和耀斑的发生情况,应用多元判别的方法建立了太阳耀斑预报模型,该模型可用于预报未来48 h内的太阳耀斑情况^[8]。

为进一步提高太阳耀斑的预报准确率,本文综合考虑了太阳黑子活动区参量、10.7 cm太阳射电流量等预报因子,提出了结合主成分分析(PCA)^[9]和支持向量机(SVM)^[10]的太阳耀斑预报模型。其中,太阳黑子活动区参量包括黑子群的McIntosh分类、活动区日面经度延伸、可见黑子数、黑子群的磁分类。利用以上预报因子,本文采用PCA-SVM方法建立了一种太阳耀斑预报模型,并将其预报结果与其他预报模型的结果进行了对比,结果验证了结合PCA-SVM的太阳耀斑预报模型是一种有效的预报模型。该模型能够嵌入国家天文台怀柔太阳观测基地的全日面磁场自动化观测系统^[11],为Ha耀斑爆发实时识别及监测系统^[12]提供先兆信息,使该系统能够实时、自动化的侦测耀斑的发生与否,并自动修改观测模式,从而更好地捕捉到太阳耀斑爆发初始时刻精细结构的演化形态数据,为太阳耀斑研究提供更多更有利的观测资料。

1 数据介绍

太阳黑子活动区参量与10.7 cm射电流量数据来自于美国国家海洋和大气管理局空间天气预报中心的每日活动区总结报告(<ftp://ftp.swpc.noaa.gov/pub/warehouse/>)。该报告记录了每日黑子活动区的观测数据,包括当日可见的所有黑子活动区的NOAA编号、位置、面积和分类情况等。我们从中直接选取了黑子群修正面积、磁分类、McIntosh分类、可见黑子数、活动区日面经度延伸等太阳黑子活动区参量。考虑到10.7 cm射电流量和太阳活动的变化密切相关^[7],本文还选取了太阳射电流量作为预报因子。X射线耀斑级别按照美国GOES卫星观测的软X射线峰值流量的量级分级,分别为A, B, C, M和X 5个等级。本文以太阳活动区为基本单位,考虑对地球影响较大的M级及以上(即M级和X级两个等级,下同)耀斑的发生与否。若一个活动区在48 h内爆发一个M级以上耀斑,则认为该活动区有耀斑发生;否则则认为没有耀斑发生。

本次模型的预报因子是基于上述数据而产生的,共有26个属性列:黑子群修正面积1个属性列,黑子群的McIntosh分类根据整体形态、最大黑子形态以及

前导和后随黑子之间分布情况等分为17个属性列(7+6+4),活动区日面经度延伸1个属性列,可见黑子数1个属性列,黑子群的磁分类4个属性列,10.7 cm射电流量1个属性列。数据分类分为+1和-1两类,+1代表该活动区48 h内会发生M级及以上耀斑,-1类代表该活动区48 h内无M级以上耀斑。除磁分类和McIntosh分类外,提取出的属性皆为实数,这些实数经过归一化处理后可直接作为预报因子的输入,而磁分类和McIntosh分类属性另作处理,具体处理过程如下。

(1) McIntosh分类。McIntosh分类第一个字母含义与Zurich分类法类似,共有7类(A, B, C, D, E, F, H)。若一个黑子群属于某一类,则其对应列赋值为1,否则赋值为0(如:若某个黑子群属于B类,则McIntosh的前7个属性赋值为0, 1, 0, 0, 0, 0, 0);McIntosh分类第二个小写字母有6种(分别为x, r, s, a, h, k),同上述规则一致,若某个黑子群属于s类,则McIntosh的中间6个属性赋值为0, 0, 1, 0, 0, 0;McIntosh分类第三个小写字母有4种(分别为x, o, i, c),同上,若某个黑子群属于i类,则McIntosh的最后4个属性赋值为0, 0, 1, 0;因此,若某个黑子群McIntosh分类为Bsi,则其属性值为:0, 1, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 1, 0。

(2) 磁分类。Wilson磁分类共有4种:单极群Alpha类,双极群Beta类,极性混杂的多极群Gamma类,同一个半影内出现两块以上不同极性的本影,其间距小于2°的Delta类。若磁分类为Beta-Gamma-Delta,则其磁分类属性值为:0, 1, 1, 1。

数据集样本范围涵盖2014年1月至12月的每一个活动区。该年数据集共有2394个样本。模型输入为经过预处理的预报因子,共包含26个属性值;模型输出为该活动区未来48 h内是否发生耀斑情况,发生耀斑则作为正例,标记为+1,不发生耀斑则作为反例,标记为-1。

2 预报模型

主成分分析(Principal Component Analysis, PCA)^[9]是一种掌握事物主要矛盾的统计分析方法,它可以从多元事物中解析出主要影响因素,揭示事物的本质,简化复杂的问题。利用转换矩阵,PCA可以将特征从高维降到低维。支持向量机(Support Vector Machine, SVM)^[10]是一种二类分类模型,其基本模型定义为特征空间上的间隔最大的线性分类器,其学

习策略便是间隔最大化,最终可转化为一个凸二次规划问题的求解。

对于线性不可分的情况,SVM通过巧妙使用核函数(kernel function),使在低维特征空间中线性不可分的样本转化为在高维特征空间其线性可分的数据,该模型的优化目标最后可写为(具体计算与转化细节见文献[10]):

$$\max_{\alpha} \left[\sum_{i=1}^m \alpha_i - \frac{1}{2} \sum_{i,j=1}^m y_i y_j \alpha_i \alpha_j k(x_i, x_j) \right]. \quad (1)$$

该支持向量机由 m 个支持向量组成,式(1)中 α 为引入的一组拉格朗日乘子, α_i, α_j 分别对应第 i 和第 j 个拉格朗日乘子, x_i, x_j 分别表示第 i 和第 j 个支持向量, y_i, y_j 分别对应 x_i, x_j 的标记值(+1或者-1),本文核函数 $k(x_i, x_j)$ 取径向基函数(Radial Basis Function, RBF)^[13],即:

$$k(x_i, x_j) = \exp(-\gamma \|x_i - x_j\|^2), \gamma > 0, \quad (2)$$

式(2)中 γ 为该径向基函数的宽度参数,控制着函数的径向作用范围, $\|x_i - x_j\|$ 表示差向量 $x_i - x_j$ 的模(2范数^[14]).在构造数据模型上,本文提取了太阳活动区的特征参量的原始数据,并利用属性编码将预报因子编码为26个输入参量,然后将这些输入参量归一化至 $[-1, +1]$ 范围内.考虑到天文数据具有数据量大的特点,也为方便接下来数据的进一步处理,本文先利用PCA法提取出主要特征,然后利用SVM对数据进行训练并分类.具体方法如下:

设Train为训练集,Test为测试集,训练集Train共有27列,其中前26列为输入的样本数据;而最后1列为标记列,代表该组样本是否有耀斑发生。

步骤1: 将训练集数据按列归一化至 $[-1, +1]$ 范围。

步骤2: 利用公式(3)和(4)得到计算数据的协方差矩阵,并计算协方差矩阵的特征值、特征向量,式(3)中 n 为样本数,协方差 $\text{cov}(x, y)$ 用于度量维度 x, y 的相关性,式(4)中 C 为各维度间的协方差构成的协方差矩阵,由于本输入样本由26个维度组成,故 m 等于26,

$$\text{cov}(x, y) = \frac{\sum_{i=1}^n (x - \bar{x})(y - \bar{y})}{n-1}, \quad (3)$$

$$C = \begin{bmatrix} \text{cov}(x_1, x_1) & \text{cov}(x_1, x_2) & \dots & \text{cov}(x_1, x_m) \\ \text{cov}(x_2, x_1) & \text{cov}(x_2, x_2) & \dots & \text{cov}(x_2, x_m) \\ \dots & \dots & \dots & \dots \\ \text{cov}(x_m, x_1) & \text{cov}(x_m, x_2) & \dots & \text{cov}(x_m, x_m) \end{bmatrix}. \quad (4)$$

步骤3: 根据计算出的特征向量,选取前 k 个主成

分构成新基, k 的选取准则见文献[15].

步骤4: 根据新基构建合适的转换矩阵,对原数据进行投影转换,得到转换后的预报因子。

步骤5: 选取RBF函数作为核函数,利用SVM训练转换后的训练集,选出合适参数对(图1为SVM训练图,该训练过程是选取合适的惩罚因子 C ^[10]与径向基参数 γ 的过程,不断尝试不同的参数对 C 与 γ ,每次计算当选用该参数对时数据分类的准确率,最后选取准确率最高的参数对作为最终训练参数对;图1横坐标为参数 C 的 $\log_2 C$ 值,纵坐标为参数 γ 的 $\log_2 \gamma$ 值,本模型选取的最终参数对为: $C=0.03125, \gamma=0.125$.各参数对的分类准确率如图1所示)。

步骤6: 将测试集数据进行归一化处理,并利用步骤4中构建的转换矩阵数据进行转换。

步骤7: 选取步骤5中得到的参数,利用SVM对步骤6中转换后的数据进行测试。

步骤8: 利用公式(5)求出分类结果,式中各参数说明见文献[10]:

$$f(x) = \text{sgn} \left(\sum y_i \alpha_i \cdot k(x, x_i) - b \right). \quad (5)$$

3 预报结果分析

数据集共2394个样本,其中发生耀斑的样本数有221个,未发生耀斑的样本数有2371个.显然,该数据集为非平衡数据集,为均衡样本中正反例比例,本文采用重采样的方法均衡数据,具体采样过程如下:随机从2173个反例样本中抽取与正例样本数目相同的样本,得到正反比例为1:1的442个样本.在实验数据验证方面,数据采用十折交叉验证,即将数

图1 (网络版彩图)SVM训练进程图

Figure 1 (Color online) The training process in SVM

数据集随机分成10份, 轮流将其中9份作为训练集, 剩下1份作为测试集. 为了保证评测指标并不是过拟合的结果, 需要进行10次实验, 并且每次都使用不同的测试, 然后将10次实验测出的评测指标的平均值作为最终的评测指标. 预报结果以混淆矩阵^[16]表示, 在二分类问题中, 如果将一个正例正确预测为正例, 则认为产生了一个真正例(true positive, TP); 如果将一个反例正确预测为反例, 则认为产生了一个真反例(true negative, TN). 相应地, 将正例预测为反例则称为伪反例(false negative, FN), 而将反例预测为正例则称为伪正例(false positive, FP).

本文与文献[17]中综合因子模型的预报结果进行了对比, 对比结果描述如下: 在预测为正例的样本中, 文献[17]的TP样本数有222例, FP样本数有63例, 两者比例约为3.52, 而PCA-SVM模型的这一比例约为4.79(139/29); 在预测为反例的样本中, 文献[17]的TN样本数(263例)与FN样本数(104例)比例约为2.53, 相较于PCA-SVM模型预测值的2.34(192/82)较高. 从这些数据可以看出, PCA-SVM模型有较好的预测正例的能力, 而这恰好是耀斑预测者希望看见的, 因为在太阳耀斑预报中, 正例的重要性显然是高于反例的. 同时, 虽然PCA-SVM模型预测反例的能力略低于文献[17], 但综合正反例数据来看, 本预报模型总的准确率仍高于文献[17]. 因此, 无论是从预测正例准确率还是从总的准确率来看, PCA-SVM模型都是一个有效的太阳耀斑预报模型.

为了更直观地描述预报结果, 本文以正例准确率(定义为 $TP/(TP+FP)$), 该指标描述了预测为正例的样本中真实发生耀斑的比例), 召回率(定义为 $TP/(TP+$

FN), 该指标描述了真实发生耀斑的样本中被准确预测到的比例)、反例准确率(定义为 $TN/(TN+FP)$)、虚报率(定义为 $FP/(FP+TP)$)4个指标查看预报结果. 文献[17]综合因子模型的上述4个指标值分别为77.9%, 68.1%, 80.7%和22.1%; 对应地, 本文PCA-SVM模型的四个指标值分别为82.7%, 62.9%, 86.9%和17.3%.

从上述4个指标的对比结果可以看出, 文献[17]模型的召回率较高, 说明该模型预测为正例的真实正例占有所有真正例的比例较高, 但另一方面, 无论是考虑正例样本还是反例样本, PCA-SVM模型都具有较高的准确率; 综合考虑所有样本数据, PCA-SVM预报模型具有更高的准确率以及更小的虚报率. 这充分验证了结合主成分分析和支持向量机的太阳耀斑预报模型是一种有效的太阳耀斑预报模型.

4 结论

本文综合考虑了太阳黑子活动区参量、10.7 cm 太阳射电流量等预报因子, 提出了结合主成分分析(PCA)和支持向量机(SVM)的太阳耀斑预报方法, 并利用此方法建立了一种有效的太阳耀斑预报模型. 该模型可预测未来48 h内是否会发生M级及以上耀斑. 对比前人太阳耀斑预报模型, 该PCA-SVM模型具有更高的准确率以及更小的虚报率. 另外, 该预报模型只涉及M级及以上耀斑的有无, 没有具体细分耀斑的级别, 且预报因子更多地考虑太阳活动区黑子群的参量. 在今后的研究中要适当考虑耀斑的具体分级以及综合考虑其他预报因子(如光球磁场特征参量等), 以便将该PCA-SVM模型推广为更广泛的太阳耀斑预报模型并进一步提高耀斑预报的准确率.

参考文献

- 1 Lin Y Z. Introduction to Solar Physics (in Chinese). Beijing: Science Press, 2000. 459–524 [林元章. 太阳物理导论. 北京: 科学出版社, 2000. 459–524]
- 2 Schuurmans C J E. Influences of solar flare particals on the general circulations of the atmosphere. *Nature*, 1965, 205: 167
- 3 Cui Y M, Li R, Zhang L, et al. Correlation between solar flare productivity and photospheric magnetic field properties. *Sol Phys*, 2006, 237: 45–59
- 4 Wang H M, Song H, Jing J, et al. The relationship between magnetic gradient and magnetic shear in fivesuper active regions producing great flares. *Chin J Astron Astrophys*, 2006, 6: 477
- 5 Li R, Cui Y M. Solar flare forecasting method of combining support vector machine with nearest neighbors (in Chinese). *Comput Eng Des*, 2009, 15: 3605–3607 [李蓉, 崔延美. 结合支持向量机和近邻法的太阳耀斑预报方法. 计算机工程与设计, 2009, 15: 3605–3607]
- 6 Gallagher P T, Moon Y J, Wang H. Active-region monitoring and flare forecasting. *Sol Phys*, 2002, 209: 171–183
- 7 McIntosh, Patrick S. The classification of sunspot groups. *Sol Phys*, 1990, 125: 251–267

- 8 Zhang G Q, Wang J L, Li D Q. A detailed introduction to the short term X-ray flare prediction scheme (in Chinese). *Prog Geophys*, 1994, 9: 48–53 [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]
- 9 Hotelling H. Analysis of complex statistical variables into principal components. *J Vision*, 1932: 417–520
- 10 Vapnik V N. The nature of statistical learning theory. *IEEE Trans Neural Networks*, 1995, 10: 988–999
- 11 Lin J B, Shen Y B, Zhu X M, et al. Huairou solar observing station full automatic observation system of surface magnetic field (in Chinese). *Astron Res Technol*, 2013, (4): 392–396 [林佳本, 沈洋斌, 朱晓明, 等. 怀柔太阳观测基地全日面磁场自动化观测系统. *天文研究与技术*, 2013, (4): 392–396]
- 12 Lin J B, Guo J, Deng Y Y. An intelligent method for solar flare observation. *Proc Int Astron Union*, 2012, 8: 503–504
- 13 Andrew A M. An introduction to support vector machines and other kernel-based learning methods. *Ai Mag*, 2000, 32: 1–28
- 14 Rugh W J. *Linear System Theory*. Upper Saddle River: Prentice Hall, 1996. 1353–1359
- 15 Jolliffe I. *Principal Component Analysis*. New York: John Wiley & Sons, Ltd, 2002. 41–64
- 16 Townsend J T. Theoretical analysis of an alphabetic confusion matrix. *Percept Psychophys*, 1971, 9: 40–50
- 17 Li R, Zhu J, Cui Y M. Solar flare forecasting model with active region photospheric magnetic field properties and sunspot factors (in Chinese). *Chin Sci Bull*, 2013, 58: 1845–1850 [李蓉, 朱杰, 崔延美. 结合活动区光球磁场参量和黑子参量的太阳耀斑预报模型. *科学通报*, 2013, 58: 1845–1850]

Solar flare forecasting model of combining principal component analysis with support vector machine

YUAN Fei^{1,2}, LIN JiaBen¹, DENG YuanYong¹, GUO JingJing¹, WANG Gang³, ZENG Zhen¹ & WANG BingXiang¹

¹ Key Laboratory of Solar Activity, Chinese Academy of Sciences, Beijing 100012, China;

² University of Chinese Academy of Sciences, Beijing 100049, China;

³ Beijing Jiaotong University, Beijing 100044, China

Solar flare is an important event in the solar atmosphere. It is a main disturbance resource of space environment and has a dramatic impact on human activities. How to promote the forecast ability of solar flare has a great significance for human beings. To better predict solar flare, this paper used principal component analysis and support vector machine, which are two classic machine learning methods, to predict solar flare. In feature extraction part, principal component analysis is a generally adopted method, and it is a technique for linearly compressing multidimensional data into lower dimensions with minimal loss of information. In flare classification part, support vector machine is a learning machine based on the statistical learning theory. The support vector machine can deal with nonlinear problems in classification and regression easily by using kernel functions, which is necessary to map onto another high dimension linear space. In solar flare forecasting research, the relationship between solar flare and morphological evolution of sunspots plays an important role in daily flare forecasting. And there is evidence that sunspot parameters and 10.7 cm solar radio flux are extremely related with solar flare. With the predictors fully taken into account, a new method of combining principal component analysis with support vector machine is proposed to improve the forecast ability of solar flare. In this paper, the sunspot parameters are area of sunspot group, McIntosh classification, extended longitude, the sunspot number in the solar active region and magnetic classification. Using attribute coding and appropriate transform function, the initial data set of predictors is normalized to form the modeling data set. And based on this data set, principal component analysis and support vector machine are applied to build solar flare forecasting model. In experiment, the forecasting model is compared with other model, which works well in the solar flare short-term forecasting. This shows that the PCA-SVM prediction model has higher precision compared with the other one when either considering positive examples or negative samples. The results indicate that the Solar flare forecasting model of combining principal component analysis with support vector machine is an effective flare prediction model.

solar flare, predictors, support vector machine, principal component analysis, sunspot

doi: 10.1360/N972015-01342