密级 _____

中国科学院大学
University of Chinese Academy of Sciences

硕士学位论文

太阳耀斑实时监测与预报方法的研究

作者姓名 袁飞

指导教师 林佳本 高级工程师

中国科学院国家天文台

学位类别 理学硕士

学科专业 天文技术与方法

培养单位 中国科学院国家天文台

2017 年 5 月

Research on the Real-time Monitoring and Forecasting Method of Solar Flare

By

Fei Yuan

A Dissertation Submitted to
University of Chinese Academy of Sciences
In partial fulfillment of the requirement
For the degree of
Master of Astronomical Technology and Method

National Astronomical Observatories, Chinese Academy of Sciences

May, 2017

摘　要

太阳耀斑的实时监测与预报不仅对太阳物理的科学研究具有重要的科研意义，而且对人类生产活动具有巨大的实用价值。为了更好地服务于太阳耀斑的实时监测与预报工作，本论文结合国家天文台怀柔太阳观测基地（HSOS）的观测数据，应用机器学习、图像处理等技术开展了太阳耀斑的监测、预报等研究，主要工作与研究成果如下：

（1）应用机器学习方法提出了一种新的太阳耀斑预报模型——结合主成分分析与支持向量机的太阳耀斑预报模型，即PCA-SVM预报模型。作为一种结合了PCA和SVM 两者优势的预报分类模型，它具有以下三点优势：① 由于前期的“降噪”和“去冗余”等处理，该分类模型在特征工程中能提取到更“精确”的特征；② 核函数（高斯核）的引入有效避免了“维数灾难”的问题且大大降低了模型训练测试过程中的运算量；③ “最大间隔”理论使该分类模型具有较好的泛化能力。

（2）将PCA-SVM预报模型应用于太阳耀斑预报的研究，并验证了PCA-SVM预报模型在太阳耀斑预报中的有效性。该方法可用于短期的太阳耀斑预报，并可以作为一种太阳活动水平评估手段。在太阳耀斑预报因子上，本文有针对性地对其进行了包括数据无量纲化、二值化、特征编码等在内的数据处理操作；在模型验证上，搜集并测试了2014 年全年共2394份样本数据，结果验证了PCA-SVM 太阳耀斑预报模型的有效性。

（3）应用引导滤波算法和OTSU阈值分割算法构造出一种新的太阳活动特征自动检测方法（简称GF-OTSU检测方法），并结合怀柔观测基地的数据，测试、验证了GF-OTSU方法的有效性。该方法充分利用了两者的优势，实现了太阳活动特征（包括太阳黑子、谱斑、耀斑等）的自动检测，应用于HSOS太阳观测数据的测试结果显示该自动检测方法与人工检测方法具有较好的一致性。

（4）在上述工作的基础上，完成了怀柔观测基地的太阳耀斑实时智能化监测系统的算法开发以及功能结构设计工作。该智能化监测系统包括“观测管理主系统”、“太阳活动水平评估模块”和“耀斑捕捉定位模块”，三者相互协同，共同服务于太阳耀斑的实时智能化监测。

关键词： 太阳活动，耀斑预报，自动监测，支持向量机，引导滤波

Abstract

The real-time monitoring and forecasting of solar flares are not only of great scientific significance for the scientific research of solar physics, but also of great practical value for social life. With the observation data of the National Astronomical Observatory of Huairou Solar Observing Station, we researched on the solar flare real-time monitoring and forecasting using machine learning and image processing technology. The main work and results are as follows:

(1) A new solar flare forecasting model is proposed, which is based on principal component analysis and support vector machine (PCA-SVM). As a combination of PCA and SVM forecasting classification model, its advantages are as follows: ① due to the previous process of "noise reduction", "data redundancy" and other processing, the classification model could extract more accurate features in Feature Engineering; ② the introduction of kernel function, such as gaussian kernel, makes it easier to avoid the problem of "Curse of Dimensionality" and greatly reduce the amount of computation in the process of training and testing; ③ the theory of "maximal margin" makes the classification model have good generalization ability.

(2) The PCA-SVM forecasting model is used on the research of the solar flares forecasting and then we verified the effectiveness of the PCA-SVM forecasting model in the solar flare forecasting. The model can be used to forecast the short-term solar flares and evaluate the level of solar activities. We have carried out the data processing operations including data non-dimensional, binary transformation and feature encoding and so on for the solar flare forecasting factor. Besides, this PCA-SVM solar flare forecasting model has been tested on the 2394 valid samples from the year of 2014, and the results show that the model is effective.

(3) We propose a new method for automatic detection of solar activities characteristics by using guided filter and OTSU algorithm (GF-OTSU detection method). And then we test and verify the effectiveness of GF-OTSU method with the HSOS data. This method embodies the advantages of guided filter and OTSU algorithm, and achieves the goal of the automatic detection of solar activities, including sunspots,

plages, flares, etc. We have tested on the data of HSOS and the results show that the method has good consistency with the manual one.

(4) On the basis of the above work, we have studied on the development of real-time intelligent monitoring system for Huairou Solar Observing Station. The intelligent monitoring system includes “Observation Management System”, “Solar Activities Level Evaluation Module” and “Flare Capture and Positioning Module”. All of them work together to serve the real-time monitoring of solar flares.

Keywords: Solar activities, Flare forecasting, Automatic monitoring, Support vector machine, Guide filter

目　　录

插　　图

表　格

第一章　绪论

1.1　本论文的课题背景

1.1.1　太阳活动及研究意义

太阳是唯一一颗可以被人类精细观测的恒星天体，是一颗基本稳定的典型恒星[1][2]。然而，大量的观测数据表明，太阳在稳定辐射的同时，也时常会有时间和空间上的局部短时活动现象，如太阳黑子、耀斑、暗条和日冕物质抛射等，这些发生在太阳大气里的活动现象统称为太阳活动[3]。

太阳黑子是在光球层中明显比背景暗黑的斑点状的小区域，是太阳表面可以看到的最突出的现象之一[4]。太阳黑子与太阳磁场密切相关，它的形成与消失通常需要几天甚至几个星期不等，长期的观测表明，当太阳黑子较多时，其他的太阳活动也会相对较频繁，并且，绝大多数的太阳活动现象发生在以黑子为核心的一个活动中心，即太阳活动区内[5]。黑子既是太阳活动区的核心，也是活动区内最明显的标志之一。研究太阳黑子不仅有利于认识太阳黑子自身变化的规律，也便于认识其他的太阳活动现象[6]。

太阳耀斑是发生在太阳大气局部区域的剧烈太阳活动现象之一，耀斑的发生常常伴随着各波段的电磁辐射以及高能粒子的辐射等现象[7][8]。当用$H\alpha$单色光监视太阳色球层时，有时会看到色球谱斑中的突然增亮现象，该局部区域的亮度增亮为原先亮度的几倍甚至几十倍，然后在数分钟至数小时内恢复至原先亮度，所以，太阳耀斑现象早期也被称为色球爆发[9]。太阳耀斑的爆发过程涉及很多复杂的物理现象，这一过程中释放的能量可以高达10^{32} erg，耀斑活动引起的电磁辐射和高能粒子辐射对地球大气气压、大气电状态等都具有明显的扰动作用，进而影响导航、通讯、广播等人类活动[10]。通常太阳耀斑的爆发可分为脉冲相、闪相和下降相三个阶段，脉冲相为耀斑开始后γ射线、硬X射线以及射电厘米波段急剧变化的阶段；而在闪相阶段，软X光、可见光以及分米波射电辐射会有几分钟的持续增强，耀斑亮度增亮几倍甚至几十倍正是在这一阶段；闪相之后，辐射缓慢减弱，这一减弱时期称为下降相[11]。在闪相阶段，耀斑亮度变化幅度巨大，现有的观测设备很难适应这一变化，也正因为这个原因，通常人们看到的耀斑影像中的耀斑都是“饱和”的，缺少具体的细

节[2]。因此，无论是为了捕捉到太阳耀斑爆发初始时刻及演化过程中的精细结构，为太阳物理的研究提供更好的观测资料，还是为了人类的生产活动，太阳耀斑的实时监测与预报研究都有着极其重要的研究意义。

1.1.2 太阳活动预报研究现状及发展趋势

太阳活动预报是空间天气预报的核心组成部分，对空间天气预报有着导向性的作用，作为一个有着多年研究历史的方向，人们在深入研究太阳活动预报的同时取得了一系列的研究成果。随着人类生产活动现代化程度的不断提高以及开发空间规模的不断扩大，太阳活动预报方法在要求一定准确度的同时有了更加多样化、明确化的需求。太阳活动预报包括周期性活动的预报和爆发性活动的预报，具体有太阳黑子、耀斑、日冕物质抛射、太阳质子事件以及太阳射电流量等的预报。根据预报时间长短的不同，太阳活动预报可分为长期预报、中期预报以及短期预报三类。这三类预报并没有严格的界限，一般认为，长期预报是指提前一年或者几年至几十年甚者更长时间的太阳活动预报，中期预报是指提前几天或者几个月的太阳活动预报，而短期预报是指提前一至三天或者几小时甚者几十分钟的太阳活动预报[12]。太阳活动预报根据预报时间尺度的不同，其预报方法各不相同。鉴于本论文的研究主题，下面将主要介绍太阳耀斑这一太阳活动现象的短期预报。目前太阳耀斑的短期预报可分为先兆法、物理预报法、经验公式法以及机器学习法等[13] [14]。

[1] 先兆法

在太阳耀斑发生之前，人们常常可以在可见光波段，射电波段或者X射线波段等观测到该区域的某些异常现象。这些现象超前于太阳耀斑的发生时间，因此，人们可以利用这些先兆现象作为太阳耀斑预报因子。目前，常用的先兆现象有黑子群特殊运动或变化，色球的暗条活动及纤维规整排列，不同层次预热现象或短波辐射增强，纵磁场中性线的变化等。根据这些先兆现象，有经验的学者可以相当准确地预报出太阳耀斑的发生。该预报方法较大地依赖于该学者的主观经验，具有较大的主观性，未来应朝着客观预报的方向上改进。

[2] 物理预报法

物理预报法是基于对耀斑储能过程的认识而逐渐发展起来的一种方法。很多学者试图通过无力磁场能量和无力因子的研究来进行太阳耀斑的预报。大量的观测数据表明，耀斑的位置与活动区速度场视向分量的反变线有关，理论上

认为活动区速度场与磁场的相互作用的观测和研究会有助于对耀斑物理过程的了解，由磁场的观测所推算出的电流核来预报耀斑可能是一种耀斑物理预报的途径。随着人们对太阳耀斑发生的物理机制理解的逐步深入，物理预报法作为一种潜在有力的预报方法正日益受到科学界的重视。

[3] 经验公式法

经验公式法使用活动区参量与耀斑产率的统计关系做预报，大量的工作在于定出统计公式。这类预报的时间提前量为一到三天，其提前量、预报水平都与导出经验公式时所依据的资料和预报时使用的资料有关。该方法具有较好的实用性，并且随着观测数据的不断增长，该方法的应用也越来越广泛。目前，经验公式法的主要分析技术包括自回归分析、模糊分析、小波分析等。

[4] 机器学习法

机器学习法（或人工智能法）是一种基于观测数据的统计预报方法，该方法可认为是一种广义上的经验公式法。与普通的统计方法不同，该方法引入了机器学习、人工智能等概念，利用现有的数据以及机器学习、人工智能的算法特性，使预报模型能够“自主”地归纳出太阳耀斑的发生规律，最终达到耀斑预报的目的。常见的机器学习的方法有人工神经网络、支持向量机、决策树和CNN算法[15]等。目前，该方法在结合合适的预报因子时具有相当的实用性，随着观测数据的不断增多以及机器学习算法的不断发展，其耀斑预报准确率将越来越高。

在太阳活动预报多年的研究历史中，国内外学者做出了不懈的努力，产出了众多的太阳耀斑预报研究成果。上世纪80年代，科罗拉多学院和美国国家海洋和大气管理局空间环境实验室联合开发了一个以太阳黑子群的McIntosh分型为主要知识基的专家系统来进行太阳耀斑的预报[16]。美国空间环境服务中心利用多元线性回归分析了McIntosh分型参数对太阳耀斑的贡献，该研究分析了McIntosh分型与太阳耀斑发生的关系[17]。大熊湖天文台根据第22 太阳活动周的历史观测数据，进一步统计分析了McIntosh分型与太阳耀斑发生的关系，并据此提出了一种基于泊松分布的太阳耀斑预报方法，该预报方法可以对各太阳活动区产生C 级、M级、或X级的X射线耀斑的可能性进行预报[18]。Leka等人通过提取合适的光球磁场数据的特征参量，利用统计方法进行了太阳耀斑的预报[19]。基于序列时间内的耀斑发生情况，Wheatland 提出了一种贝叶斯太阳耀斑预报方法，该方法可用于全日面GOES事件的预报[20]。近些年来，一些学

者使用横向磁场强度、纵向磁场强度、磁场梯度变化和磁剪切等参量，结合统计学习方法进行太阳耀斑的预报并取得了较好的实验结果 [21] [22] 。

在国内，国家天文台太阳活动预报组利用多元回归方法建立了太阳耀斑预报方法，该预报方法可预报未来48小时内M级以上的太阳耀斑，涉及的预报因子包括：(1)太阳活动区光球纵向场数据；(2)包括太阳黑子群面积、磁分类和McIntosh 分型的太阳黑子光学观测数据；(3) 10.7cm射电流量数据 [23] 。该方法在很长一段时间作为太阳活动预报中心每日预报的主要参考工具，之后，Zhu 等人采用该方法用于实际的太阳活动预报并与美国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 [24] 。另外，国内学者在应用机器学习方法预报太阳耀斑上也做了一系列的工作：选取上述主要的预报因子，Li等人应用支持向量机和k近邻方法建立了一个耀斑短期预报系统，并得到良好的预报精度 [25] 。Wang等人将从光球磁图中提取的纵向磁场最大水平梯度、中性线长度和孤立奇点个数等参量，应用于多层感知器建立起了一个太阳耀斑短期预报系统 [26] 。Huang等人应用几种机器学习方法建立了耀斑预报模型，结果显示了磁场数据时间序列演化信息在耀斑预报中的有效性 [27] 。采用神经网络和聚类相结合的耀斑预报方法，Li等人的研究结果表明了聚类算法在耀斑预报建模中的重要性 [28] 。位于国家天文台怀柔太阳观测基地的太阳磁场望远镜，从1987年开始常规观测，现已平稳运行三十年，Yang等人利用该磁场望远镜二十余年的矢量磁场数据，从统计角度分析了太阳光球活动区磁场非势性强度随太阳活动周演化的关系，然后利用一种简单的通用机器学习方法检验了太阳耀斑预报方法中磁非势性参量的预报性能 [29] 。

近些年来太阳耀斑预报技术取得了很大的发展，虽然限于目前人们对耀斑爆发物理机制理解的局限性，完全基于物理模型的太阳耀斑预报模型仍然较少，但是，随着该领域研究的逐步深入，预报因子参量的选择将越来越精准与细致，基于物理模型的太阳耀斑预报能力将不断提高。另一方面，随着计算机技术的不断发展，新的太阳耀斑预报方法也不断出现，这些方法将具有更快的运算速度与更准确的预报率。未来一段时间内的主要发展趋势仍然是通过不断加深对耀斑爆发物理机制的理解、开发更加有效的算法模型，结合其他太阳活动特征的监测情况，提高综合预报的能力。

1.1.3　相关望远镜简介

中国科学院国家天文台怀柔太阳观测基地创建于1984年，是国家天文台的重要观测基地之一，它是太阳物理界享有高知名度的太阳磁场和速度场观测台站与学术研究中心。经过三十年的发展，怀柔观测基地目前拥有多台套高分辨率、高灵敏度的局部磁场和全日面磁场望远镜等多台套太阳望远镜系统，主要包括 60cm 三通道太阳望远镜 [30] 、35cm 太阳磁场望远镜 [31] 、$H\alpha$色球望远镜以及全日面磁场望远镜 [32] 。

[1] 60cm 三通道太阳望远镜

怀柔观测基地配备的 60cm 三通道太阳望远镜是真空格里高利式反射望远镜，该望远镜滤光器选择 5173Å、5247Å 和5250Å 三条工作谱线，可在同一时刻获得三个不同层次上的太阳磁场数据，为描绘出一幅立体的太阳磁场结构图提供更加有力的工具。天文工作人员可以通过该望远镜的观测资料进行太阳光球层到色球层的磁场演变分析等研究。

[2] 35cm 太阳磁场望远镜

该望远镜曾是世界上最先进的太阳磁场观测设备之一，它能够获得λ = 532.419 nm 的光球和λ = 486.134 nm 色球的矢量磁场以及视线速度场数据，有效视场为4.1′ × 3.5′，CCD大小为1K × 1K，像元分辨率约0.24″/pixel。自投入使用以来，该望远镜取得了大量国际一流的观测数据，为太阳物理的研究提供了宝贵的观测资料。作为一台多功能的综合性太阳观测望远镜，它可用于太阳磁场、磁活动的观测研究以及具有空间灾害天气的监测与预报等。

[3] $H\alpha$色球望远镜

怀柔观测基地 $H\alpha$色球望远镜装备有$H\alpha$ 滤光器以及拍摄耀斑及其他突发现象的面阵 CCD 摄像机。主要参数有：口径 20 cm，有效焦距 180 cm，配备半宽 0.025nm 的Lyot双折射滤光器，滤光器放在准直光路中，探测器有效像元数 2712×2712 ，像元分辨率为 0.86339 ″/pixel。目前，该全日面色球望远镜已与其他四个台站组成国际联测网，可为一些太阳活动（尤其是耀斑）的观测提供珍贵的观测资料。本文的耀斑监测、预报等研究工作就是在该望远镜上开展的。

[4] 全日面磁场望远镜

怀柔观测基地的全日面磁场望远镜配备半宽 0.01nm 的双折射滤光器，可以在532.4 nm 处观测太阳窄带单色像和矢量磁场，探测器像元数 992×992 ，常规观测Stokes V/I、Q/I、U/I 的谱线位置为偏离线心 -0.008nm ，观测视场约 33′

，像元分辨率约 1.93″/pixel 。该望远镜可用于观测全日面矢量磁场、纵速度场以及相应的光学图像。本文的太阳活动特征的检测等相关研究工作是在该望远镜上开展的。

1.2 本论文的研究目的及内容

在太阳耀斑的监测与预报中，既有基于太阳物理知识、由人工来实现的监测与预报手段，也有基于计算机技术等、由机器自动实现的手段。目前，国内外现有的技术方法或运算复杂度较高，不利于耀斑的实时监测；或采用了一些较复杂的预报因子，不利于怀柔基地现有观测系统的特征提取；并且，作为通用的监测预报技术，其很难完全符合怀柔基地现有的条件基础。因此，发展出一套符合怀柔基地现实情况的太阳耀斑监测预报系统显得尤为必要了。本课题的研究目的是依托于国家天文台怀柔太阳观测基地前期发展积累下来的理论、技术基础以及现有的观测设备，通过机器学习、图像处理等技术手段，开展太阳耀斑的监测、预报等研究工作，以提高怀柔基地太阳耀斑实时监测与高分辨观测的水平和能力。主要内容可分为“太阳耀斑的预报”、“太阳活动特征的检测”和“太阳耀斑的实时智能化监测”三部分。“太阳耀斑的预报”和“太阳活动特征的检测”可为后期的“太阳耀斑的实时智能化监测”提供必要的理论与技术基础，而“太阳耀斑的实时智能化监测”是“太阳耀斑的预报”和“太阳活动特征的检测”的进一步实践，三者相互协同，共同服务于怀柔基地的太阳耀斑实时智能化监测。

[1] 太阳耀斑的预报

机器学习是一门涉及统计分析、概率论和线性代数等多领域的综合性学科 [33] ，也是一种智能化的、能够随着训练次数的增加而性能不断提升和优化的学习系统，该系统能够赋予计算机“学习能力”并使其在没有明确编程的情况下做出合适的反应 [34] 。受益于机器学习的这一优势，太阳耀斑预报研究又有了一些新的方法。本文利用机器学习技术开展了太阳耀斑的分类、预报等研究工作。在这一部分工作中，本文围绕耀斑预报因子和耀斑预报方法两个方面开展了研究。在耀斑预报因子上，本文研究的预报因子主要包括太阳黑子活动参量以及射电流量数据等，对于不同的预报因子，本文有针对性地进行了特定的数据处理方法；在耀斑预报方法上，提出了一种具有较强泛化能力的PCA-SVM预报模型，并利用2014年全年的样本数据进行测试，结果验证了

该预报模型的有效性。

[2] 太阳活动特征的检测

单一地依靠文本数据并不能很好地服务于太阳耀斑的智能化监测，往往需要一些其他太阳活动特征的检测结果。在基于前文“太阳耀斑预报”的基础上，开展了基于图像的“太阳活动特征检测”的工作。在这一部分工作中，本文首次将引导滤波这一图像处理技术引入到天文图像的目标检测中，并提出了一种可应用于怀柔基地现有设备的太阳活动特征检测方法。该检测方法可用于检测太阳黑子、谱斑、耀斑等太阳活动特征。怀柔基地的测试结果表明，该太阳活动特征检测方法的检测结果与人工检测结果具有较好的一致性。

[3] 太阳耀斑的实时智能化监测

在前期工作的基础上，本文将上述预报、检测技术嵌入至怀柔基地现有的观测系统，在测试阶段已能实现太阳耀斑的实时智能化判别，并具有实时修改观测模式的功能，可为太阳耀斑的研究提供更多更有利的观测资料。实验表明，无论是基于文本的预报，还是基于图像的检测，这些技术都是功能有效的。并且，相较于单一考虑文本或图像的监测系统，本课题太阳耀斑监测系统的选用方法和参数具有一定的互补优势。

1.3 本论文的结构安排

鉴于本课题为服务于怀柔太阳观测基地的应用型研究课题，本论文将紧紧围绕“太阳耀斑的实时智能化监测”这一主题展开。太阳耀斑的预报可以作为一种评估太阳活动水平的手段，太阳活动特征的检测可应用于太阳耀斑的捕捉定位，无论是“太阳活动水平评估”还是“耀斑捕捉定位”，都将最终服务于怀柔太阳观测基地的太阳耀斑实时智能化监测系统。在结构布局上，本文可分为五个章节。除本章的引言部分，第二章以机器学习方法为技术背景，以耀斑为研究对象，详细阐述了基于机器学习技术的太阳耀斑预报方法的研究。第三章介绍了以图像处理技术为基础的太阳活动特征方法的研究，然后以太阳谱斑、太阳黑子为例对这一特征检测方法进行了描述。在介绍太阳耀斑预报、太阳活动特征检测方法研究的同时，本文也一并介绍了包括机器学习、图像处理技术在内的相关算法与技术。第四章节在前文的基础上阐述了太阳耀斑实时智能化监测系统的研制。最后，第五章是对本文的总结以及相关研究的展望。

第二章　太阳耀斑预报方法研究

2.1　方法概述

2.1.1　机器学习概述

机器学习（Machine Learning，ML）是最近二三十年来较为活跃的一门多领域交叉学科，涉及统计分析、概率论和线性代数等多门学科，是一种智能化的，能够随着训练次数的增加而在性能上不断提升和优化的学习系统 [35] [36]。作为计算机科学的一支分枝，机器学习同时也是一种能够赋予计算机学习能力并让计算机在没有明确编程情况下做出合适反应的科学。通俗地说，机器学习是一种应用数据训练模型，最后使用训练出的模型进行聚类、分类和预测等应用的方法 [37]。人类不断成长的过程就是一个不断“学习”的过程：在学习成长的过程中逐渐积累了很多的历史经验，然后通过“归纳”、“总结”这些经验获得生活的“规律”。于是，当人遇见未知的问题或事物时，能够根据原先获知的规律去对这一未知问题或事物进行合理的“推测”，从而指导自己的生活。举个简单的例子，人类能够认出一棵以前从未见过的树为“树”，就在于人类能够进行“归纳”、“推测”，进而理解一些“规律”。与此类似，机器学习“训练模型”、“预测分类”的过程可对应于人类学习过程中“归纳”和“推测”的过程。通过这样的一个学习过程，计算机可以在没有具体的编程（或者说没有具体的因果逻辑）的情况下，“学习”到一些很复杂的事物，即从数据中挖掘潜在的规律，“学习”到一些有价值的事物 [38]。

机器学习算法有很多，根据“学习”任务的不同可分为监督学习和无监督学习 [39]，监督学习和非监督学习的最大差别在于机器学习的训练集中是否有正确值的标记，有则为监督学习，没有则为非监督学习。在具体算法上，监督学习主要可分为统计分类 [40]、回归分析 [41] 等；无监督学习主要有聚类 [42] 和关联规则 [43] 等。

机器学习现今已经有了十分广泛的应用，如生物特征识别、数据挖掘、垃圾邮件分类、信用卡欺诈检测和语音识别等 [44]。除此之外，机器学习在天文技术中的应用也同样十分广泛。它不仅能够帮助自动地检测某些太阳活动特征，也能够利用训练好的模型在某些尚未明确物理机制的太阳活动中做出一些

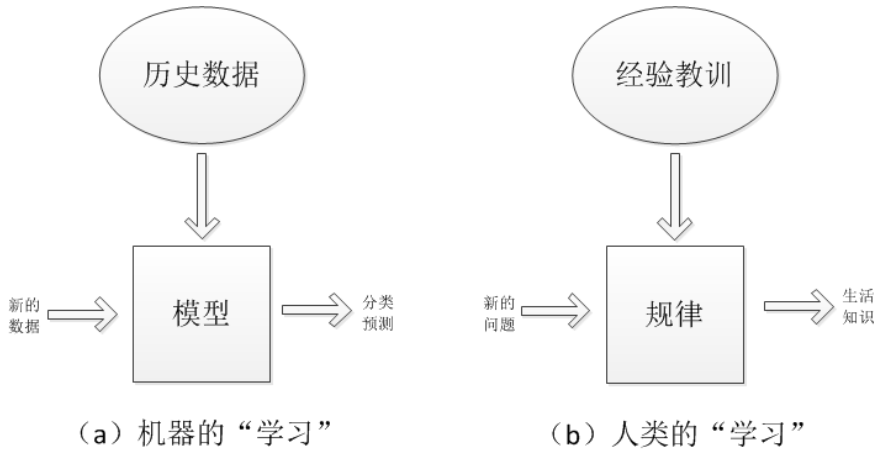


图 2.1: 机器学习与人类学习的类比

合理的预测，比如太阳耀斑的预测等 [45] 。机器学习遍布于人类生活的方方面面，随着计算机技术以及电子元器件的不断发展和提升，机器学习正在不断改变着人们处理事务的方式。

2.1.2 分类预报算法

机器学习中用于处理分类预报等问题的算法较多，例如决策树、贝叶斯分类算法、人工神经网络、KNN算法、支持向量机等；另外也有由多个分类算法组合而成的学习算法，如AdaBoost、随机深林和Boosting 等 [46] 。而对于多分类问题，可以由多个单分类器组合而成。

2.1.2.1 决策树

作为经典分类算法之一，决策树 [47] 可用于数据的分类与预测。它是一种树状预测模型，每一棵决策树都由唯一的一个根节点、多个内部节点以及作为路径终点的叶子节点组成，树的结构如图 2.2 所示。根节点是路径的始端，是全体训练数据、测试数据的集合，内部节点可以看成一个该树的分支，它将一个大的问题分为多个小问题。决策树的分类过程以根节点出发节点，不断根据所在节点的特征分类进入下一个节点，直至达到叶子节点，做出最终的分类决策。图 2.2 是预测小明是否会出去玩的决策树，利用这棵树，可以对新增的记录进行分类，判断出在其他的条件下小明是否会外出游玩。从这棵决策树可以看出，当“没有空”时，小明对于“会不会出去玩”选择了“No”，而当“有

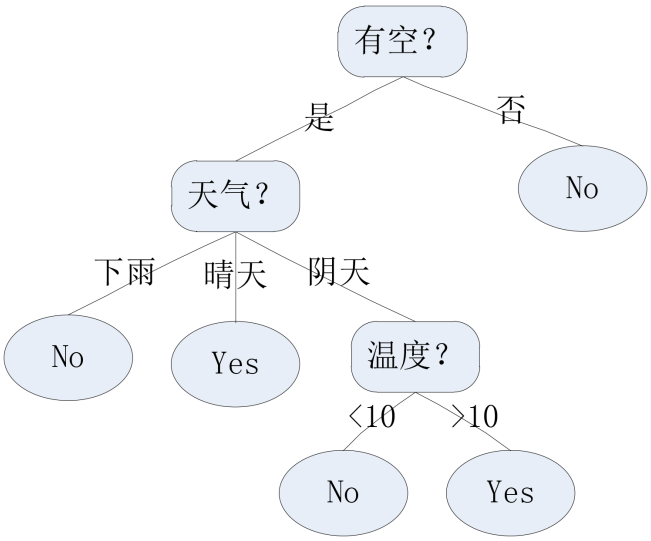


图 2.2: 一棵典型的决策树

空”时，则需进一步考虑其他条件，比如天气、温度等，直至最终到达叶子节点，即确定“是否出去游玩”。决策树可以较为直观地表示最终的分类、预测结果。

决策树用于分类的基本步骤为：（1）创建合适的数据集；（2）计算信息熵、信息增益、信息增益率等，将数据集进行划分，创建决策树；（3）根据创建的决策树进行分类。决策树分类的关键在于树的构建。常见的决策树算法有ID3 和C4.5，这两种算法是分别根据信息增益和信息增益率来作为分类依据的。决策树的创建过程如图 2.3 所示。

创建树的过程中，可以根据遍历所有特征后的信息增益来决定分裂节点的特征。假设在样本数据集 D 中，有 n 种类别数据。根据选择的特征，计算这些数据的信息熵 $Info(D)$：

$$Info(D) = -\sum_{i=1}^{c} p_i * log_2(p_i) \tag{2.1}$$

对于有c个类别的数据集D，p_i 为第i类样本占数据集D样本数的比例，$Info(D)$ 为数据集D的信息熵。

当选择特征X作为分裂节点时，特征X作用后的数据集D的信息熵为

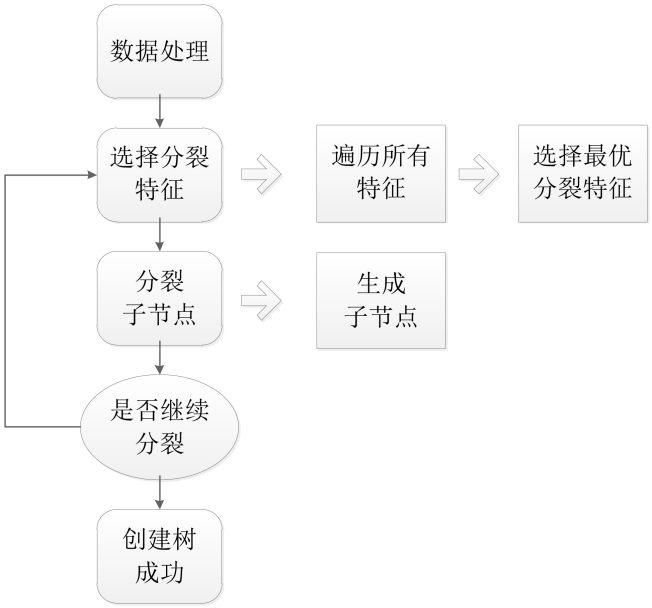


图 2.3: 决策树的创建流程

$Info_X(D)$，如公式 2.2 所示：

$$Info_X(D) = -\sum_{j=1}^{k} \frac{|D_j|}{|D|} * Info(D_j) \tag{2.2}$$

在公式 2.2 中，数据集D根据特征X被分为了k个部分，$Info(D_j)$表示第j类样本的信息熵，$\frac{|D_j|}{|D|}$ 表示第j类样本占整个数据集D的比例。信息增益被定义为数据集在选择某个特征后信息熵减少的值，其公式如下：

$$Gain(X) = Info(D) - Info_X(D) \tag{2.3}$$

决策树在创建过程中将选择使得信息增益$Gain(X)$最大的特征X 作为分裂特征。决策树在分类过程中不仅有着计算复杂度不高、输出的结果也较为直观（树形结构）、便于算法使用者的理解等优势，同时，它也具有对于中间值缺失不很敏感、对数据的独立性要求不高等优点。然而，决策树存在剪枝难以把握，可能会出现欠拟合或过拟合的问题。

2.1.2.2 贝叶斯分类算法

贝叶斯分类算法（Bayes）[48] 是一类基于贝叶斯定理预测样本类别的分类算法的总称。对于一个未知样本，该分类算法将选择其中概率最高的类别作为

判定的类别。贝叶斯定理是在独立性的假设前提下才成立的，而对于工程应用中的数据，这一假设条件是几乎不可能达到的。为此，工程应用中出现了很多对独立性假设要求较低的贝叶斯分类算法，朴素贝叶斯分类算法则是贝叶斯分类算法中最为简单的一种。

对于一个给定的样本 x ，该样本属于 y 类的概率为：

$$P(y|x) = \frac{P(x|y) * P(y)}{P(x)} \tag{2.4}$$

其中，P(x)、P(y)分别为x、y的先验概率，$P(x|y)$表示在y类别中x出现的概率。如果x样本特征向量为$\vec{x}$，基于各特征相互独立的假设上，则样本x属于c_k类的概率为：

$$P(y = c_k|x) = \frac{\prod_{i=1}^{M} P(x^i|y = c_k) * P(y = c_k)}{\sum_k P(y = c_k) \prod_{i=1}^{M} P(x^i|y = c_k)} \tag{2.5}$$

其中，数据集共有k个类别，$P(y = c_k)$表示在所有样本中，某个样本属于第k个类别的概率。向量x由M列组成，$P(x^i|y = c_k)$ 表示在类别c_k中向量x的第i个维度出现的概率。因此，可以通过公式 2.5 计算出每一个样本出现的条件概率。

朴素贝叶斯分类算法的基本流程如下（图 2.4 ）：

（1）若某个样本x为一个待分类项，它由M个维度（特征）组成，即x=$a_1, a_2, ..., a_M$；

（2）数据集共有k个类别，C=$y_1, y_2, ..., y_k$；

（3）计算条件概率$P(y_1|x), P(y_2|x), ...P(y_k|x)$；

（4）比较步骤（3）中的各条件概率，若最大条件概率为$P(y_k|x)$，则认为样本x属于第k类。

朴素贝叶斯作为一种经典的分类算法，它具有如下几点优势：1、算法原理简单，需要估计的参数很少，对缺失数据不是很敏感；2、支持增量式的运算，可以实时快速地训练新增的样本；3、分类器的结果易于理解。但是，因为独立性的假设原因，朴素贝叶斯分类算法的缺点也十分明显：当特征个数相对较多或者各个特征之间具有较大的相关性时，该分类器的分类准确率将大大降低。

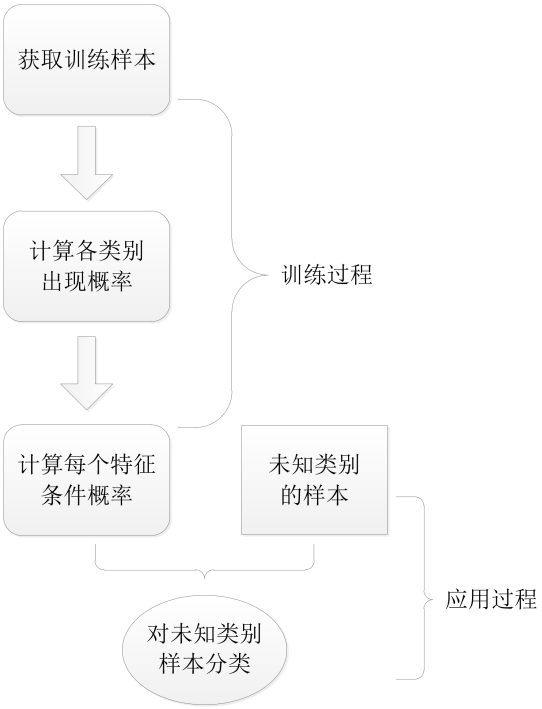


图 2.4: 朴素贝叶斯分类流程

2.1.2.3 人工神经网络

人工神经网络（Artificial Neural Networks，ANN）[49] 是一种模拟人类大脑传递信息的数学模型。在这一模型里，大量的网络节点（神经元）相互连接，进行信息传递。神经网络通常需要大量的训练才能最终达到“学习”的目的。人工神经网络进行“学习”的过程就是在多次的训练过程中不断更新各网络节点的权值的过程，最后，训练好的神经网络模型可以用于样本的分类。如图 2.5 所示，人工神经网络通常由三层相同的层状网络构成，即输入层、中间层以及最后的输出层。

目前神经网络已有上百种模型，其中，常见的有前馈神经网络（Feed-forward Neural Network）[50]、后反馈神经网络（Back-propagation Network）[51]、随机神经网络（Stochastic Neural Network）[52] 和竞争神经网络（Competitive Neural Netwo）[53] 等。通常人们所说的人工神经网络大多为后反馈神经网络。人工神经网络作为分类算法具有的优点有：1、有一定的联想能力，能逼近任意非线性关系。2、有较强的容错性。3、具有较强的学习能力，在某些场景下

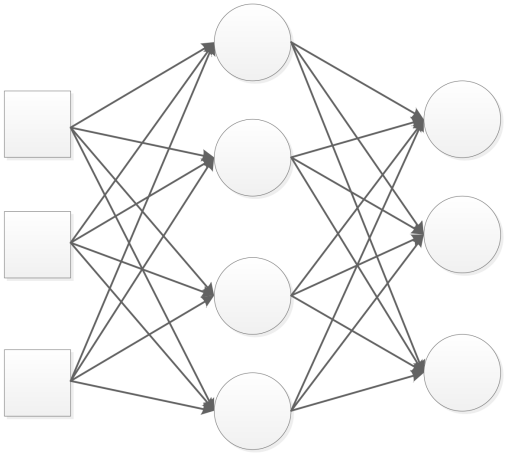


图 2.5: 人工神经网络示意图

分类准确率极高。同时，它具有的缺点也很明显，包括：1、神经网络参数较多，权值和阈值等不好把握。2、中间层是隐藏的，不能直接观察中间的结果。3、学习过程相对较长，有可能陷入某个局部极小值。神经网络在深度学习领域也具有很大的影响力，目前深度神经网络常常被用于自然语言处理、计算机视觉、语音或文字识别（图 2.6 ）等领域，并取得很好的效果。

2.1.2.4　KNN算法

K-近邻（K-Nearest Neighbors，KNN）[54] 分类算法是一种基于原有数据实例的分类算法，它的核心思想十分简单，即通过计算不同特征间的距离，找出离该样本最近的那K个邻居，然后将该样本划分至这K个邻居所体现（即最多的那一个类别）的那一类上。

对于如图 2.7 所示的样本分类过程中，若K为3，离未知圆形样本最近的3个样本中有2个三角形，1 个正方形，根据KNN的思想，则该未知样本应归为“三角形”类别；但若K为9时，此时，指定区域内有5 个正方形和4个三角形，则未知样本应归于“正方形”类。

KNN算法有着类似于“物以类聚，人以群分”的指导思想，由最近邻的 K 个已知样本来确定对象的分类。它的基本流程可以概括为：（1）计算距离：对于给定样本，计算它与已知样本中各个对象的距离或相似度；（2）找出K 个最

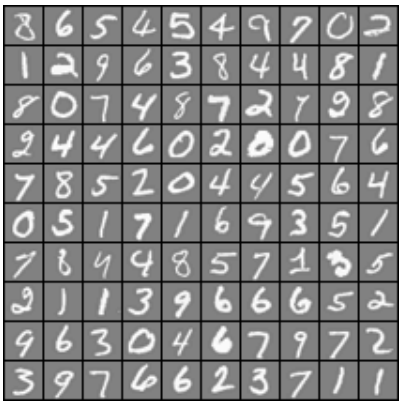


图 2.6: 人工神经网络在文字识别上的应用

近的邻居：由上一步的计算结果，找出给定样本最近的K个邻居训练样本；（3）确定分类：执行“少数服从多数”的原则，K个邻居中最多的那一类别作为给定样本的最终分类结果。

在KNN分类算法中，距离或相似度的计算是极其关键的一个环节。在机器学习中，距离或相似度的度量主要有欧式距离（Euclidean Distance）[55]、曼哈顿距离（Manhattan Distance）[56]、切比雪夫距离（Chebyshev Distance）[57]和余弦相似度（Cosine Similarity）[58]等。

[1] 欧式距离

欧式距离源于欧几里得几何，也是大众最为熟知的一种距离计算方式。在二维空间中，两点间的欧氏距离为：

$$d_{ab} = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} \tag{2.6}$$

其中，a、b两点坐标分别为(x_1, y_1)和(x_2, y_2)。推广到n维向量，则向量$\vec{a}$和$\vec{b}$距离为:

$$d_{\vec{a}\vec{b}} = \sqrt{(\vec{a} - \vec{b})(\vec{a} - \vec{b})^T} \tag{2.7}$$

向量$(\vec{a} - \vec{b})^T$为向量$(\vec{a} - \vec{b})$的转置。将向量展开，则两向量间的距离可以表示为：

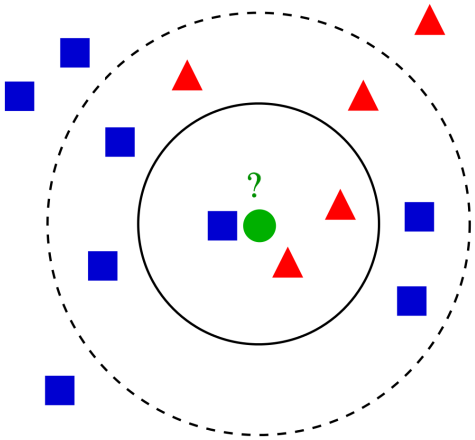


图 2.7: KNN算法的分类思想

$$d_{\vec{a}\vec{b}} = \sqrt{\sum_{k=1}^{n}(x_{1k} - x_{2k})^2} \tag{2.8}$$

n维向量$\vec{a}$和$\vec{b}$分别为$a(x_{11}, x_{12}, ..., x_{1n})$和$b(x_{21}, x_{22}, ..., x_{2n})$。

[2] 曼哈顿距离

形象地说，曼哈顿距离就像是从曼哈顿的一个十字路口走到另外一个十字路口，两点间的距离并不是一条直线，而是类似于“城市街区的距离”。在二维空间中，平面上两点$a(x_1, y_1)$和$b(x_2, y_2)$ 的曼哈顿距离为：

$$d_{ab} = |(x_2 - x_1)| + |(y_2 - y_1)| \tag{2.9}$$

而对于n维向量，$\vec{a}(x_{11}, x_{12}, ..., x_{1n})$和$\vec{b}(x_{21}, x_{22}, ..., x_{2n})$的曼哈顿距离为：

$$d_{\vec{a}\vec{b}} = \sum_{k=1}^{n}|x_{1k} - x_{2k}| \tag{2.10}$$

式中，“$|*|$”为对“$*$”取绝对值。

[3] 切比雪夫距离

切比雪夫距离与曼哈顿距离比较类似，但它的距离并不是取“城市街区的距离”，而是取两维度间最大的那一个距离，即$max(|(x_2 - x_1)|, |(y_2 - y_1)|)$。在二

维平面上，$a(x_1, y_1)$和$b(x_2, y_2)$两点的切比雪夫距离为：

$$d_{ab} = max(|(x_2 - x_1)|, |(y_2 - y_1)|) \tag{2.11}$$

而对于n维向量，$\vec{a}(x_{11}, x_{12}, ..., x_{1n})$和$\vec{b}(x_{21}, x_{22}, ..., x_{2n})$的切比雪夫距离为：

$$d_{\vec{a}\vec{b}} = max_k(|x_{1k} - x_{2k}|) \tag{2.12}$$

[4] 余弦相似度

余弦夹角度量的是两个向量方向上的差异，在KNN分类算法中，余弦相似度则用于度量两个样本向量间的差异。在二维平面上两点$a(x_1, y_1)$ 和$b(x_2, y_2)$的余弦相似度为：

$$cos(\theta) = \frac{x_1 x_2 + y_1 y_2}{\sqrt{{x_1}^2 + {y_1}^2}\sqrt{{x_2}^2 + {y_2}^2}} \tag{2.13}$$

对于n维样本，$\vec{a}(x_{11}, x_{12}, ..., x_{1n})$和$\vec{b}(x_{21}, x_{22}, ..., x_{2n})$的余弦相似度为：

$$cos(\theta) = \frac{\vec{a} \cdot \vec{b}}{|\vec{a}|\,|\vec{a}|} \tag{2.14}$$

式中“·”为两个向量间的点乘。

总的来说，KNN分类算法具有算法简单、对异常数据不敏感、无数据输入的假定等优点。同时也具有K 值不好把握以及计算空间较大等缺点。

2.2 数据处理与方案设计

2.2.1 数据介绍

太阳耀斑作为一种剧烈的太阳活动爆发现象，它的发生与多种因素有关[59]。研究表明，太阳耀斑的发生与否与太阳黑子等相关变量有着密切的关系[18]。其中，太阳黑子相关参量包括磁分类、黑子群的McIntoch 分类和Zurich分类等参量[60] [14]。另一方面，磁场参量对太阳耀斑的研究具有重要意义，这些参量包括磁场剪切、磁螺度和电流螺度、纵向磁场最大水平梯度、中性线长度、孤立奇点等参量[61]。这些参量都可以作为机器学习算法模型的输入特征，即本论文中的太阳耀斑预报因子。但考虑到获取数据的有效性、连续性等因素，本文首先选取了部分的太阳黑子活动区参量以及10.7cm 射电流量等参数作为此次太阳耀斑预报方法的候选预报因子[62]。

在这些候选预报因子中，太阳活动区参量来自于太阳活动监测网页（solarmonitor.org），该网页提供近乎实时的太阳活动区、太阳活动事件的监测数据文档，含太阳图像数据（图 2.8 ）以及文本报告（图 2.9 ），该报告包含太阳活动区NOAA 编号、最近的日面位置、Hale分类、McIntosh分类、黑子群面积（当日面积和前一日面积）、黑子群内的黑子个数以及近期的耀斑发生情况。其他太阳活动区参量来自于美国国家海洋和大气管理局空间天气预报中心（SWPC）的活动区每日报告（网址为ftp.swpc.noaa.gov/pub/warehouse）。该文件报告主要描述当日太阳活动情况，包括各活动区的活动区编号、日面经纬度、黑子群面积、黑子群磁分类等。10.7cm射电流量数据来源于NOAA的太阳与地球物理活动报告（Report of Solar and Geophysical Activity，RSGA），该报告含有每日的10.7cm 太阳射电流量数据。

本文考虑的数据集样本涵盖2014年1月至12月的每一个活动区。在训练集中，若该活动区未来48 h内发生耀斑则该样本作为正例，标记为1，不发生耀斑则作为反例，标记为0。表 2.1 为2014年的太阳活动参量的样本数据（限于篇幅，这里只显示了2014年前5日的数据），表中各列分别为观测日期、NOAA编号、日面维度、日面经度、卡林顿经度、黑子群面积、McIntosh 分类、延伸经度、黑子群个数、磁分类、可见黑子数、当日射电流量和两日内耀斑发生情况[63] 。

2.2.2　特征工程

上节得到的预报因子数据是未经处理的特征数据，这些特征可能存在信息冗余 [64]、信息利用率低等问题，需要经过一定的数据预处理等操作，以便于进一步的太阳耀斑预报研究。在太阳耀斑预报等具体的工程应用中，获取到的数据往往并不能直接用作“预报因子”。这些原始数据往往需要经过一定的“加工”、“提取”和“优化”等处理后才能真正用于后期的算实践。在机器学习中，这些“加工”、“提取”和“优化”的过程被称之为特征工程 [65] 。

特征工程是机器学习过程中的一个重要组成部分。在具体的工程应用中，人们获取的数据很少能直接用于某个机器学习算法。而特征工程就是应用相关的知识来创建使机器学习算法可以更好地处理某些“特定特征”的过程。对于机器学习的整个过程而言，“特征”的好坏对于最终的学习效果具有举足轻重的影响。工程应用中大部分的机器学习模型都可以通过构造适合该数据的“特

表 2.1: 太阳活动区参量示例

Day	Num	SunLat	SunLng	KaqLng	Area	McIntosh	LngL	SpotGrp	MagType	MNum	Flux	Type
2014/1/1	1934	S16	W83	272	80	Eac	11	5	$\beta\gamma\delta$	8	160	0
2014/1/1	1936	S16	W36	225	280	Eac	14	36	$\beta\gamma\delta$	8	160	1
2014/1/1	1937	S12	W09	198	10	Bxo	4	3	β	8	160	0
2014/1/1	1938	S09	E09	179	30	Hax	2	3	α	8	160	0
2014/1/1	1940	S12	W44	233	20	Dro	4	4	β	8	160	0
2014/1/1	1941	S13	W22	211	30	Dro	4	3	β	8	160	0
2014/1/1	1942	N10	E62	127	20	Hrx	1	1	α	8	160	0
2014/1/1	1943	S11	E67	122	20	Cro	1	1	β	8	160	0
2014/1/2	1936	S16	W49	225	210	Eac	13	18	$\beta\gamma\delta$	7	161	1
2014/1/2	1938	S12	W03	178	30	Cro	4	6	β	7	161	0
2014/1/2	1940	S12	W59	235	40	Cao	4	3	$\beta\gamma$	7	161	0
2014/1/2	1941	S12	W34	210	40	Dao	5	4	$\beta\gamma$	7	161	0
2014/1/2	1942	N10	E49	126	20	Cro	1	1	β	7	161	0
2014/1/2	1943	S11	E57	118	20	Cro	3	1	β	7	161	0
2014/1/2	1944	S07	E75	101	250	Dko	7	3	β	7	161	1
2014/1/3	1936	S17	W62	225	160	Eai	13	9	$\beta\gamma$	8	182	1
2014/1/3	1938	S14	W16	179	30	Cao	5	4	β	8	182	0
2014/1/3	1940	S12	W73	236	70	Dao	6	7	β	8	182	0
2014/1/3	1941	S13	W48	211	30	Dao	5	4	β	8	182	0
2014/1/3	1942	N10	E37	126	30	Cao	6	4	β	8	182	0
2014/1/3	1943	S11	E45	118	20	Cro	3	2	β	8	182	0
2014/1/3	1944	S08	E64	99	1060	Fkc	16	19	$\beta\gamma$	8	182	1
2014/1/3	1945	N12	E18	145	10	Bxo	4	4	β	8	182	0
2014/1/4	1936	S16	W75	225	110	Dai	8	5	$\beta\gamma$	9	215	1
2014/1/4	1937	S12	W47	197	30	Dao	5	8	$\beta\gamma$	9	215	0
2014/1/4	1938	S14	W31	181	10	Bxo	3	3	β	9	215	0
2014/1/4	1940	S12	W85	235	70	Dao	5	4	β	9	215	0
2014/1/4	1941	S13	W61	211	20	Cso	5	3	β	9	215	0
2014/1/4	1942	N10	E23	127	30	Cao	5	4	β	9	215	0
2014/1/4	1943	S12	E30	120	10	Cro	2	3	β	9	215	0
2014/1/4	1944	S09	E53	97	1280	Fkc	16	37	$\beta\gamma$	9	215	1
2014/1/4	1945	N11	E04	146	10	Dro	5	5	β	9	215	0
2014/1/5	1936	S15	W90	227	60	Hsx	3	2	α	9	218	0
2014/1/5	1937	S12	W58	195	80	Dac	6	14	$\beta\gamma$	9	218	0
2014/1/5	1938	S14	W45	182	10	Hrx	2	3	α	9	218	0
2014/1/5	1941	S13	W78	215	20	Hrx	1	1	α	9	218	0
2014/1/5	1942	N10	E08	129	30	Hsx	2	1	α	9	218	0
2014/1/5	1943	S11	E17	120	20	Hrx	1	3	α	9	218	0
2014/1/5	1944	S08	E38	99	1470	Fkc	20	60	$\beta\gamma\delta$	9	218	1
2014/1/5	1945	N11	W12	149	10	Axx	1	2	α	9	218	0
2014/1/5	1946	N12	E41	96	10	Bxo	3	2	β	9	218	1

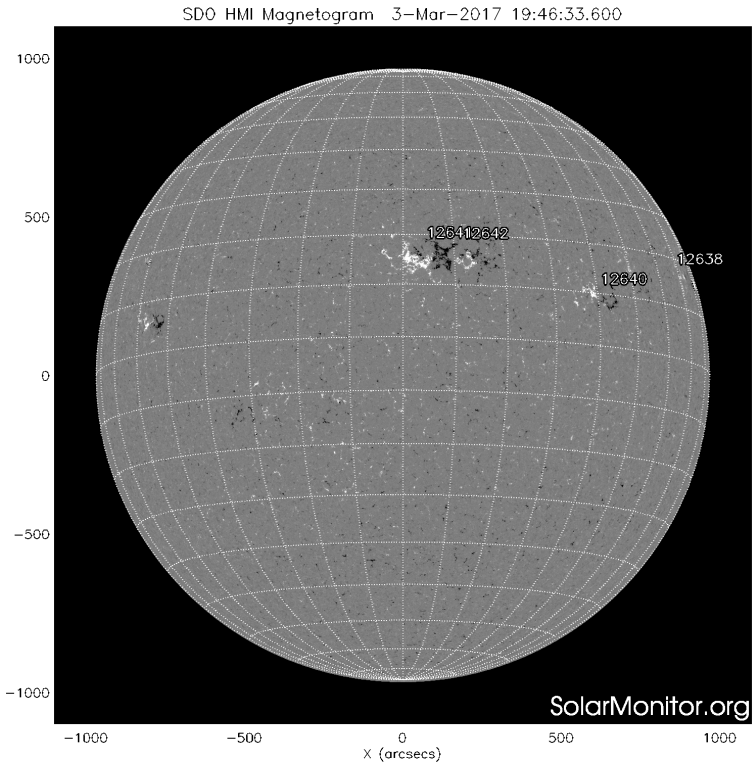


图 2.8: 太阳活动区监测图像

征”来好好“学习”，良好的特征更易得到好的识别、预测效果，也更加易于使用者理解、维护。因此，特征工程是机器学习算法流程中极其重要的一个环节。

在机器学习的学习过程中，很多原始数据（如图像数据、音频数据以及部分文本数据）可能具有成千上万的“特征”，若直接用于建模并训练，则数据量太多且不利于一些程序的“快速响应”。而特征选取则是选择原始数据中一部分合适的特征进行后期的“模型训练”，在选取的过程中也起到了“绛维”的作用。同样，在太阳耀斑的监测与预报过程中，对于数据特征较多、数据量较大且需快速响应的工程应用中，分类器需要自动地选择出对太阳耀斑预报有意义的特征进行特征选择。一般来说，可以从监测、预报特征的发散性以及这些特征与耀斑发生情况的相关性来进行选择。一个特征是否发散关系到这个特征对于样本的区分度是否有用，若特征发散度趋近于0 （即方差接近0），说明这个太阳耀斑预报因子的样本在这个特征上几乎没有差异，显然这样的特征对耀斑的监测和预报没有任何实际价值。而特征的相关性则关系到该特征与目标

NOAA Number	Latest Position	Hale Class	McIntosh Class	Sunspot Area [millionths]	Number of Spots	Recent Flares
12638	N16W81 (918",282")	α/α	Hsx/Hsx	0080/0080	01/01	-
12640	N08W45 (679",219")	α/α	Axx/Axx	0010/0010	01/02	-
12641	N15W08 (130",366")	β/β	Bxo/Dro	0030/0060	07/07	-
12642	N15W15 (242",363")	β/β	Bxo/Cro	0020/0040	03/05	-

图 2.9: 太阳活动区监测报告

相关度的大小，所以在太阳耀斑预报研究中，对于相关度高的特征，可以优先考虑。不同的处理方式对最终模型的准确度的影响不同，一些与要解决问题相关度不大的特征被认为是冗余信息，应该被移除。特征选择算法的结果可以通过测试集中模型训练好坏来进行判断。

常见的数据绛维方法[66]有主成分分析（Principle component analysis，PCA）[67] [68]、自组织特征映射（self-organizing feature map）[69]和多维缩放（Multi-dimensional scaling）[70]等。本论文的数据绛维方法主要为主成分分析。通俗地说，主成分分析就是把原有的多个数据指标转化成少数几个代表性相对较好的综合性指标，并且，这少数几个指标能够反映原先数据的大部分信息（如95%以上）。主成分分析的过程主要起着降维和简化数据结构的作用，即“降噪”和“去冗余”的作用。“降噪”使保留下来的特征的相关性尽可能地小，而“去冗余”则使得保留下来的特征含有的“能量”（方差）尽可能地大。PCA的基本过程可以分为以下四个步骤：（1）首先，形成样本矩阵，样本中心化；（2）然后，计算样本矩阵的协方差矩阵；（3）接着对协方差矩阵进行特征值分解，选取最大的 N 个特征值对应的特征向量组成投影矩阵；（4）最后，对原始样本矩阵进行投影，得到降维后的新样本矩阵。

在太阳耀斑预报方法研究中，本文的特征工程主要有以下四点。首先，对所有的预报因子进行了无量纲化处理，以确保耀斑预报算法不会过于依赖某一预报因子；然后，对于定量特征、定性特征，有针对性地对其二值化或进行特有的编码处理；最后，对于缺失的数据，采取相应的缺失值填充处理。

[1] 无量纲化

为了避免机器学习算法的预测结果过于依赖某一特征，经常需要在预处理阶段对数据进行无量纲化。无量纲化将不同尺度的数据转换至同一尺度。经典的无量纲化方法有数据标准化以及区间缩放法两种。数据标准化的前提是预报

表 2.2: 预报因子定量特征统计

	黑子群面积	黑子群个数	射电流量
均值	133.2	8.1	149.0
均值的标准误差	4.9	0.2	0.5
中值	60.0	4.0	149.0
众数	10.0	1.0	152.0
标准差	241.6	10.8	26.7
方差	58377.8	117.4	712.8
全距	2750.0	115.0	151.0
极小值	0.0	1.0	86.0
极大值	2750.0	116.0	237.0
和	318840.0	19503.0	356695.0

因子的特征服从或者近似服从正态分布，经过标准化后，预报因子则服从标准正态分布。以2014年的数据为例，本文对黑子群面积、黑子群个数和10.7cm射电流量三个特征进行了统计分析。对于这些数据，计算了包括均值、中值、众数、标准差、方差、极大值和极小值。统计数据如表 2.2 所示。另一方面，本文分别统计了这三个定量特征出现的频率，即2014年全年该变量出现的次数，图 2.10 为黑子群面积的频率统计图、图 2.11 为黑子群个数的频率统计图、图 2.12 为射电流量的频率统计图。

同时，本文分别计算了黑子群面积、黑子群个数以及10.7cm射电流量的正态P-P图 [71] （图 2.14 ），从图中可以看出，只有射电流量近似服从正态分布（该特征的趋降正态P-P图如图 2.15 所示，其正态偏差在0.03左右），拟合曲线如图 2.13 所示,该正态分布均值为149.00，标准差为26.7。

除了10.7cm射电流量这一特征，本章节其余预报因子皆为非均衡数据，并不服从（或近似服从）正态分布，因此，可用区间缩放法对这些预报因子进行无量纲化，使其归一化至[0,1]范围内。缩放方法如公式 2.15 所示：

$$x' = \frac{x - Min}{Max - Min} \tag{2.15}$$

其中，x为转换前参量，Min、Max分别为在所有记录中x参量的最小值和最大值，x'为转换后的参量。

[2] 特征二值化

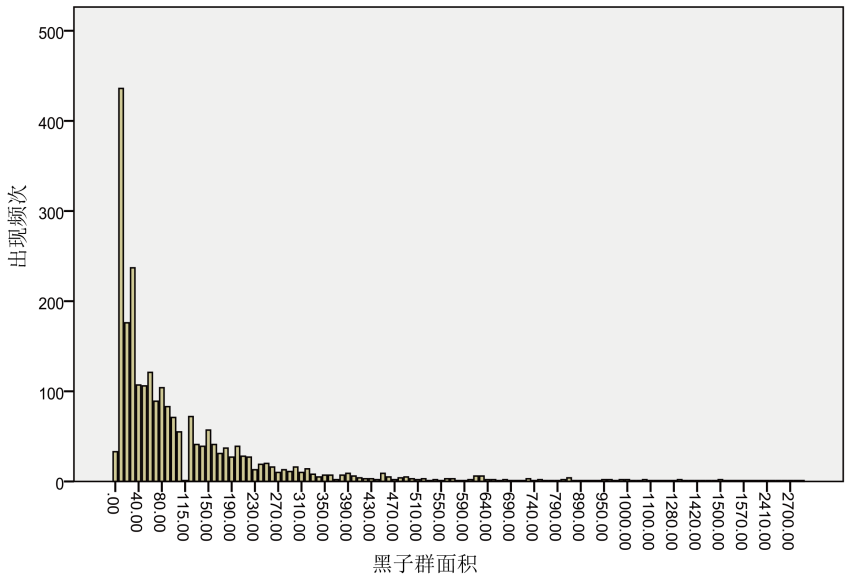


图 2.10: 黑子群面积的频率统计图

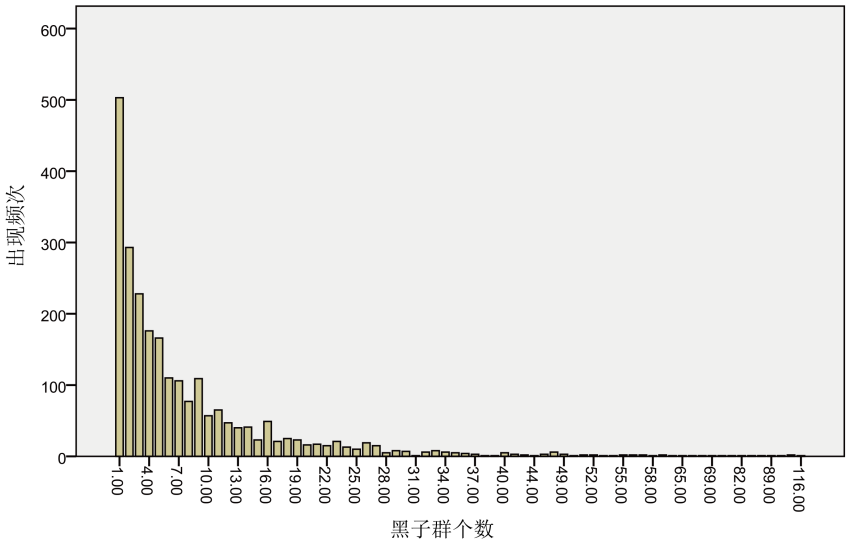


图 2.11: 黑子群个数的频率统计图

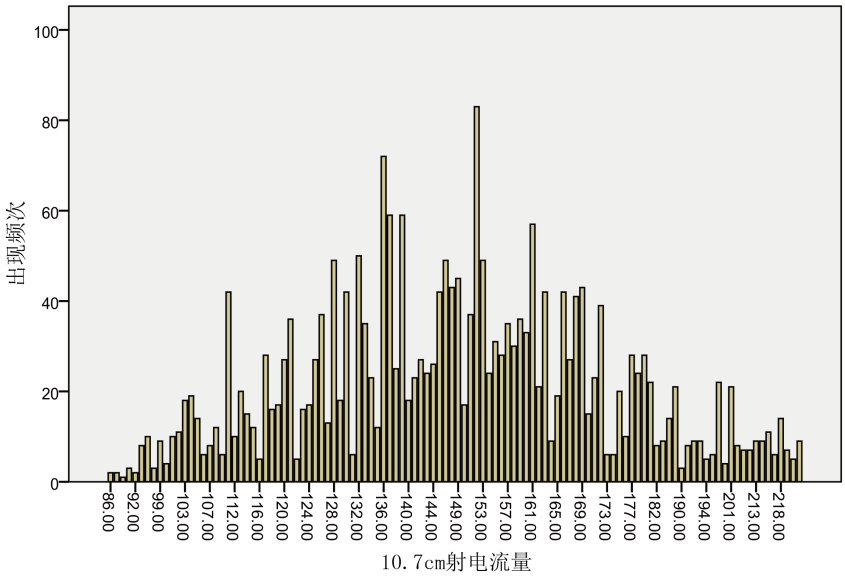


图 2.12: 射电流量的频率统计图

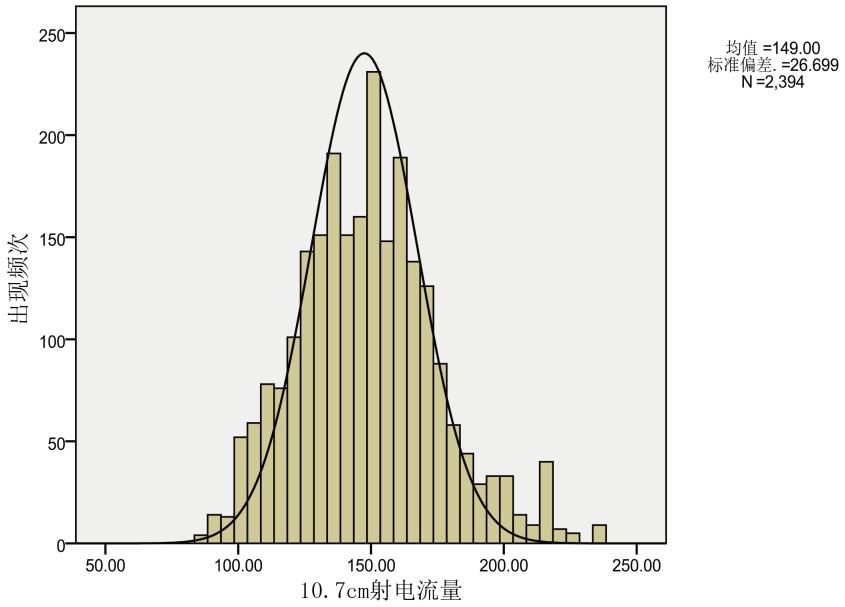


图 2.13: 射电流量频率的拟合图

特征二值化即将某个定量特征二值化处理，它的核心在于先设定一个阈值T，大于等于该阈值的赋值为 1 ，而小于该阈值的则赋值为 0 ，表达式如公式 2.16 所示：

$$x' = \frac{x - Min}{Max - Min} \tag{2.16}$$

在本次的太阳耀斑预报中，耀斑的发生情况这一特征就是经过了定量特征二值化处理：将未来两天内发生了M级及以上耀斑的太阳活动区标记为1，将未发生M级以上耀斑的活动区标记为0。

[3] One-Hot编码

对于McIntosh分类、磁分类这种离散且非连续的特征，若直接用数字表示，则表达效率会提高很多，但是，即使转化为数字后，这些特征也并不能直接用于太阳耀斑的预报分类器中。因为分类器往往默认特征数据是连续且有序的，人为加上这种有序性会给太阳耀斑预报分类器一定的误导作用。为了解决这一问题，可以使用One-Hot编码技术，其方法是将N个状态用N位状态寄存器保存，每个状态都有其独立的寄存位，并且在任意时刻只有一位数字有效[72]。对于一个特征，如果它有N个可能值，那么经过One-Hot编码后，它将转化为N个二值特征（只有0 和1两个状态），且这些特征互斥，对于每个样本数据，只有一个被激活（赋值为1），显然，经过One-Hot 编码后，数据变为稀疏[73]的了。

以McIntosh分类为例，对McIntosh分类的处理过程如下：McIntosh分类由三个字母构成，第一个字母与Zurich分类类似，共有七类（A，B，C，D，E，F，H），若一个黑子群属于某一类，则对应位赋值为1，否则为0（如：若某黑子群属于C 类，则One-Hot 编码后McIntosh分类的前七个特征分别赋值为0，0，1，0，0，0，0）；McIntosh分类的第二个字母有x，r，s，a，h，k 共6 种，若某个黑子群属于h类，则One-Hot编码后McIntosh分类的中间六个属性赋值为0,0,0,0,1,0 ；McIntosh分类的第三个字母有x，o，i，c 4 种，若某个黑子群属于x 类，则One-Hot编码后McIntosh 分类的后四个属性为1，0，0，0。因此，若某样本的McIntosh分类为Fkc，则One-Hot编码后该样本属性被赋值为0，0，0，0，0，1，0，0，0，0，0，0，1，0，0，0，1 。

One-Hot编码的好处在于，一方面它解决了分类器不好处理特征数据的问题，另一方面，它在一定程度上又扩充了特征，丰富了样本数据。

[4] 缺失值填充

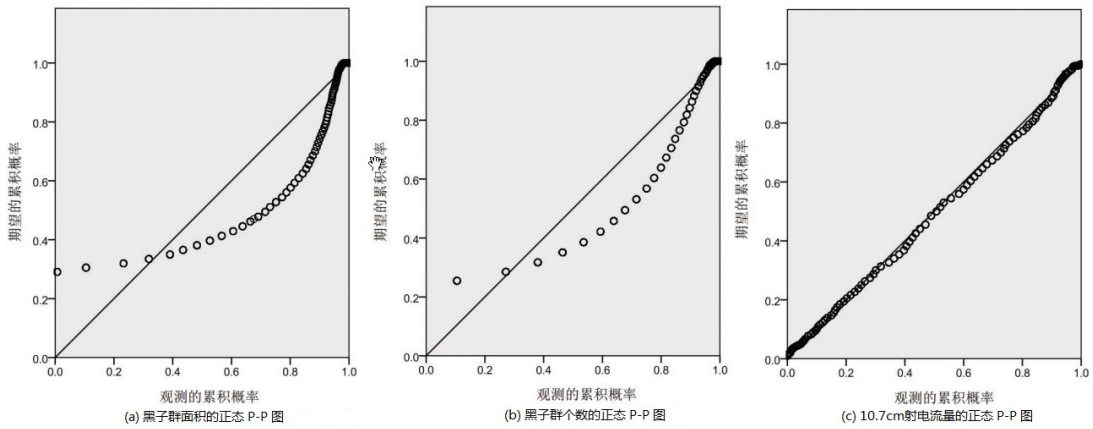


图 2.14: 预报因子中定量特征的正态P-P图

由于在数据搜集的过程中，获取的数据并不总是完整的，常常需要对缺失值进行填充。常见的缺失值填充算法有均值法 [74]、线性回归 [75]、KNN [76]、健壮贝叶斯估计（RBE） [77]等。这些算法增加数据完备性的同时，也使分类器能够更加充分地利用已经搜集到的数据。在本次太阳耀斑预报中，太阳活动区参量极少存在数据缺失的情况，故直接采用均值填充缺失值。

2014全年的太阳耀斑参量数据（部分数据如表 2.1 所示）经过无量纲化、特征二值化、One-Hot编码和缺失值填充等处理后，耀斑预报因子特征由原先11个特征转换为现今的26个特征，处理后的数据如表 2.3 所示（同样地，这里只显示了前5天的数据），表 2.3 各列依次为日期、卡林顿经度、黑子群面积、McIntosh分类的17 个特征、延伸经度、黑子群个数、磁分类的4个特征和当日射电流量和两日内耀斑发生情况。

2.2.3　特征选择

经过前期的特征工程处理后，需要选择出对太阳耀斑预报有意义的特征进行特征选择。以各预报因子对耀斑发生情况的贡献率来考虑：图 2.16 、 2.17 、 2.18 分别展示了太阳黑子群面积、黑子群个数、10.7cm射电流量对是否发生M级以上耀斑的贡献率，从图中可以看出，这三个特征对是否发生耀斑并没有太明显的趋势。

对于太阳耀斑预报这一特殊的工程应用，单一地选择某一个特征或者某几个特征并不能很好地预测太阳耀斑的发生情况，必须综合地考量各个特征性质

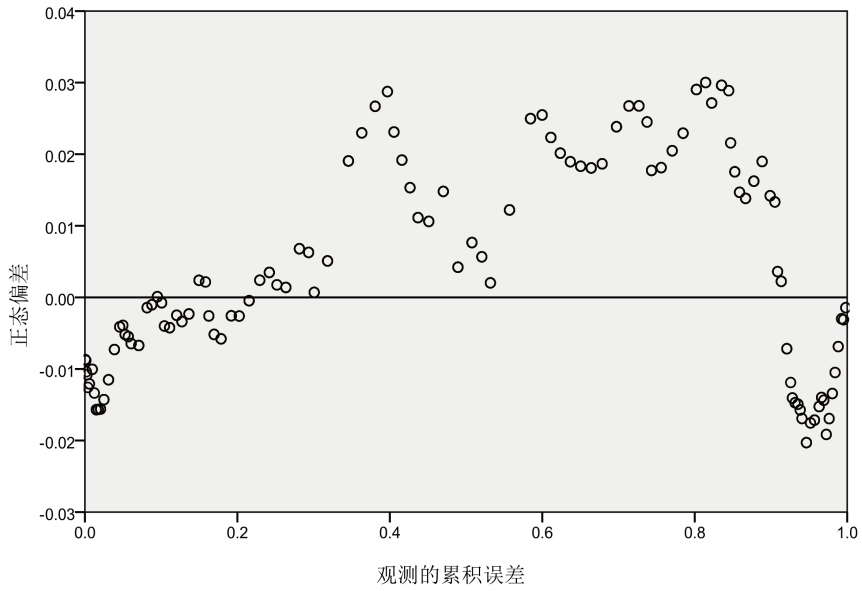


图 2.15: 10.7cm射电流量的趋降正态P-P图

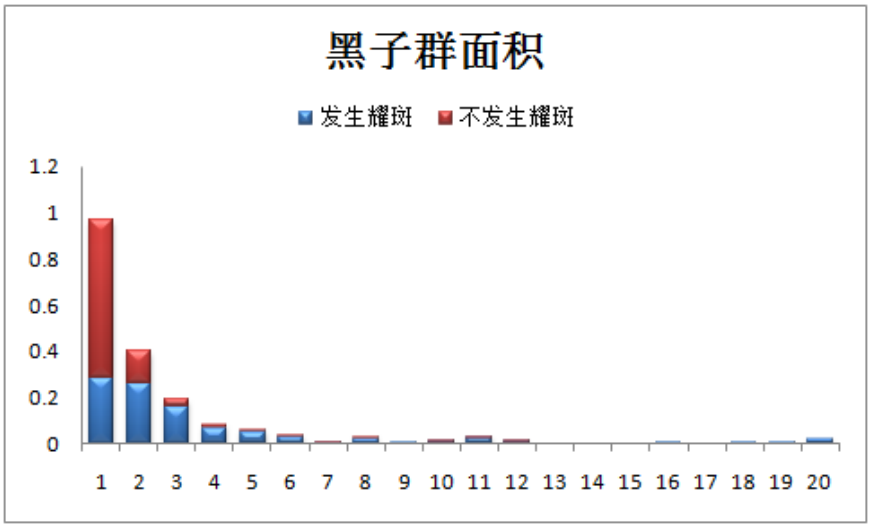


图 2.16: 黑子群面积的贡献率

表 2.3: 预处理后的太阳活动区参量示例

Day	KaqLng	Area	McIntosh 1-17	LngL	SpotGrp	MagType 1-4	Flux	Type
2014/1/1	0.756	0.029	0,0,0,0,1,0,0,0,0,0,1,0,0,0,0,0,1	0.407	0.035	0,1,1,1	0.49	0
2014/1/1	0.625	0.102	0,0,0,0,1,0,0,0,0,0,1,0,0,0,0,0,1	0.519	0.304	0,1,1,1	0.49	1
2014/1/1	0.55	0.004	0,1,0,0,0,0,0,1,0,0,0,0,0,0,1,0,0	0.148	0.017	0,1,0,0	0.49	0
2014/1/1	0.497	0.011	0,0,0,0,0,0,1,0,0,0,1,0,0,1,0,0,0	0.074	0.017	1,0,0,0	0.49	0
2014/1/1	0.647	0.007	0,0,0,1,0,0,0,0,1,0,0,0,0,0,1,0,0	0.148	0.026	0,1,0,0	0.49	0
2014/1/1	0.586	0.011	0,0,0,1,0,0,0,0,1,0,0,0,0,0,1,0,0	0.148	0.017	0,1,0,0	0.49	0
2014/1/1	0.353	0.007	0,0,0,0,0,0,1,0,1,0,0,0,0,1,0,0,0	0.037	0	1,0,0,0	0.49	0
2014/1/1	0.339	0.007	0,0,1,0,0,0,0,0,1,0,0,0,0,0,1,0,0	0.037	0	0,1,0,0	0.49	0
2014/1/2	0.625	0.076	0,0,0,0,1,0,0,0,0,0,1,0,0,0,0,0,1	0.481	0.148	0,1,1,1	0.497	1
2014/1/2	0.494	0.011	0,0,1,0,0,0,0,0,1,0,0,0,0,0,1,0,0	0.148	0.043	0,1,0,0	0.497	0
2014/1/2	0.653	0.015	0,0,1,0,0,0,0,0,0,0,1,0,0,0,1,0,0	0.148	0.017	0,1,1,0	0.497	0
2014/1/2	0.583	0.015	0,0,0,1,0,0,0,0,0,0,1,0,0,0,1,0,0	0.185	0.026	0,1,1,0	0.497	0
2014/1/2	0.35	0.007	0,0,1,0,0,0,0,0,1,0,0,0,0,0,1,0,0	0.037	0	0,1,0,0	0.497	0
2014/1/2	0.328	0.007	0,0,1,0,0,0,0,0,1,0,0,0,0,0,1,0,0	0.111	0	0,1,0,0	0.497	0
2014/1/2	0.281	0.091	0,0,0,1,0,0,0,0,0,0,0,1,0,1,0,0	0.259	0.017	0,1,0,0	0.497	1
2014/1/3	0.625	0.058	0,0,0,0,1,0,0,0,0,0,1,0,0,0,0,1,0	0.481	0.07	0,1,1,0	0.636	1
2014/1/3	0.497	0.011	0,0,1,0,0,0,0,0,0,0,1,0,0,0,1,0,0	0.185	0.026	0,1,0,0	0.636	0
2014/1/3	0.656	0.025	0,0,0,1,0,0,0,0,0,0,1,0,0,0,1,0,0	0.222	0.052	0,1,0,0	0.636	0
2014/1/3	0.586	0.011	0,0,0,1,0,0,0,0,0,0,1,0,0,0,1,0,0	0.185	0.026	0,1,0,0	0.636	0
2014/1/3	0.35	0.011	0,0,1,0,0,0,0,0,0,0,1,0,0,0,1,0,0	0.222	0.026	0,1,0,0	0.636	0
2014/1/3	0.328	0.007	0,0,1,0,0,0,0,0,1,0,0,0,0,0,1,0,0	0.111	0.009	0,1,0,0	0.636	0
2014/1/3	0.275	0.385	0,0,0,0,0,1,0,0,0,0,0,0,1,0,0,0,1	0.593	0.157	0,1,1,0	0.636	1
2014/1/3	0.403	0.004	0,1,0,0,0,0,0,1,0,0,0,0,0,0,1,0,0	0.148	0.026	0,1,0,0	0.636	0
2014/1/4	0.625	0.04	0,0,0,1,0,0,0,0,0,0,1,0,0,0,0,1,0	0.296	0.035	0,1,1,0	0.854	1
2014/1/4	0.547	0.011	0,0,0,1,0,0,0,0,0,0,1,0,0,0,1,0,0	0.185	0.061	0,1,1,0	0.854	0
2014/1/4	0.503	0.004	0,1,0,0,0,0,0,1,0,0,0,0,0,0,1,0,0	0.111	0.017	0,1,0,0	0.854	0
2014/1/4	0.653	0.025	0,0,0,1,0,0,0,0,0,0,1,0,0,0,1,0,0	0.185	0.026	0,1,0,0	0.854	0
2014/1/4	0.586	0.007	0,0,1,0,0,0,0,0,0,1,0,0,0,0,1,0,0	0.185	0.017	0,1,0,0	0.854	0
2014/1/4	0.353	0.011	0,0,1,0,0,0,0,0,0,0,1,0,0,0,1,0,0	0.185	0.026	0,1,0,0	0.854	0
2014/1/4	0.333	0.004	0,0,1,0,0,0,0,0,1,0,0,0,0,0,1,0,0	0.074	0.017	0,1,0,0	0.854	0
2014/1/4	0.269	0.465	0,0,0,0,0,1,0,0,0,0,0,0,1,0,0,0,1	0.593	0.313	0,1,1,0	0.854	1
2014/1/4	0.406	0.004	0,0,0,1,0,0,0,0,1,0,0,0,0,0,1,0,0	0.185	0.035	0,1,0,0	0.854	0
2014/1/5	0.631	0.022	0,0,0,0,0,0,1,0,0,1,0,0,0,1,0,0,0	0.111	0.009	1,0,0,0	0.874	0
2014/1/5	0.542	0.029	0,0,0,1,0,0,0,0,0,0,1,0,0,0,0,0,1	0.222	0.113	0,1,1,0	0.874	0
2014/1/5	0.506	0.004	0,0,0,0,0,0,1,0,1,0,0,0,0,1,0,0,0	0.074	0.017	1,0,0,0	0.874	0
2014/1/5	0.597	0.007	0,0,0,0,0,0,1,0,1,0,0,0,0,1,0,0,0	0.037	0	1,0,0,0	0.874	0
2014/1/5	0.358	0.011	0,0,0,0,0,0,1,0,0,1,0,0,0,1,0,0,0	0.074	0	1,0,0,0	0.874	0
2014/1/5	0.333	0.007	0,0,0,0,0,0,1,0,1,0,0,0,0,1,0,0,0	0.037	0.017	1,0,0,0	0.874	0
2014/1/5	0.275	0.535	0,0,0,0,0,1,0,0,0,0,0,0,1,0,0,0,1	0.741	0.513	0,1,1,1	0.874	1
2014/1/5	0.414	0.004	1,0,0,0,0,0,0,1,0,0,0,0,0,1,0,0,0	0.037	0.009	1,0,0,0	0.874	0
2014/1/5	0.267	0.004	0,1,0,0,0,0,0,1,0,0,0,0,0,0,1,0,0	0.111	0.009	0,1,0,0	0.874	1

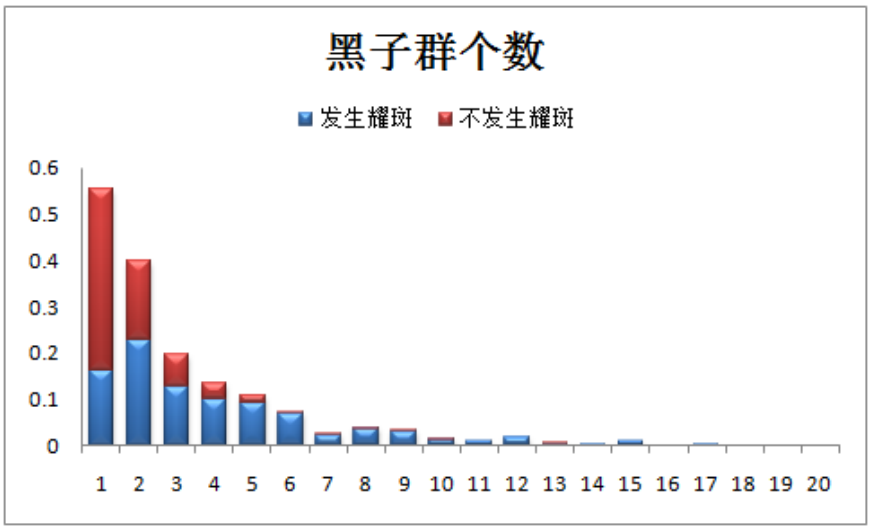


图 2.17: 黑子群个数的贡献率

和特点，选用合适的机器学习算法对其进行分类、预测。

2.2.4 太阳耀斑预报模型

2.2.4.1 支持向量机

支持向量机（Support Vector Machine，SVM）[78] 是20世纪90年代由Vapnik等人基于统计学习理论而发展起来的一种经典的机器学习分类算法。支持向量机通过结构化最小风险来最大化地提高其泛化能力，使分类超平面间隔最大化来达到良好的分类效果。对于如图 2.19 所示的线性可分的二分类样本，存在多条可能将训练样本(○,×)分开的分类线。在图 2.19 （a）中，显然分类线 a 的分类效果最好，因为相较于其他分类线，它更远离每一类样本，风险最小。而其他的分类线b、c离分类样本较近，当样本发生较小变化时将可能导致错误的分类。因此，分类线a 是代表对样本(○,×)的一个最优的线性分类器。在支持向量机中最优分类线就是使两类的分类间隔最大且将各类别准确分开的分类线。同样，在图 2.19 （b）中，分类线H1 和H2 分别为过各类样本的支持向量，H为最优分类线，支持向量H1以及支持向量H2之间的距离叫做这两个类别间的分类间隔，支持向量机算法的训练过程就是找出使分类间隔最大的分类线。对于非线性问题，支持向量机采用核函数（kernel function）把样本数据映射到一个更高维度的特征空间中，然后在这一高维空间进行线性分类，最终达到处理

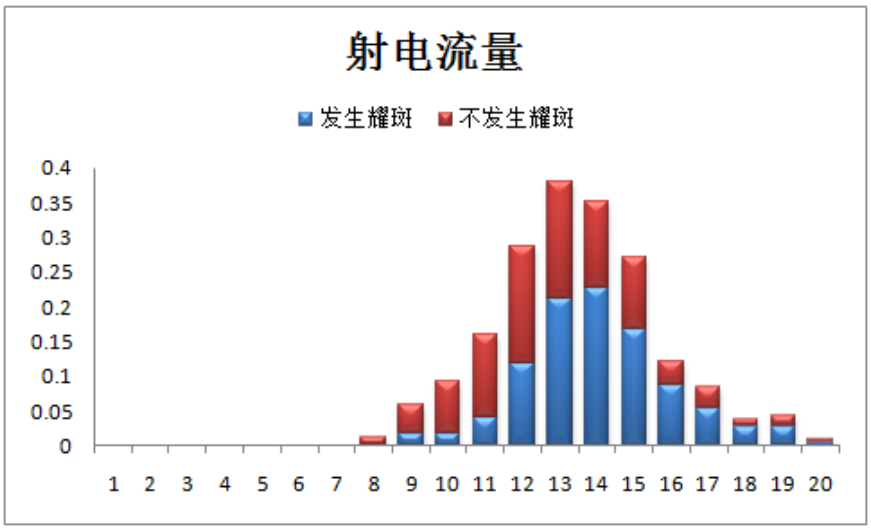


图 2.18: 10.7cm射电流量的贡献率

非线性问题的目的（图 2.20 ）。

支持向量机分类算法流程如下：

（1）给出一组输入样本x_i，i=1,2,⋯,n及其对应的期望输出yi∈+1,-1；

（2）在约束条件（公式 2.17 ）下使得分类间隔最大，使公式 2.18 的函数值最大化：

$$\sum_{i=1}^{n} a_i y_i = 0, \quad 0 \le a_i \le C \tag{2.17}$$

其中C为松弛变量相关常数，保证了支持向量机在处理线性或非线性问题时有一定的抗干扰能力。

$$Q(a) = \sum_{i=1}^{n} a_i - \frac{1}{2} \sum_{i,j=1}^{n} a_i a_j y_i y_j K(x_i \cdot x_j) \tag{2.18}$$

$K(x_i, x_j)$为核函数，它将向量x_i和x_j投影到一个更高的向量空间。

（3）计算向量的权值：

$$W^* = \sum_{i=1}^{n} a_i^* y_i x_i \tag{2.19}$$

$$b^* = \frac{1}{y_s} - W^* \cdot x_s \tag{2.20}$$

其中 x_s 为一个特定的支持向量。

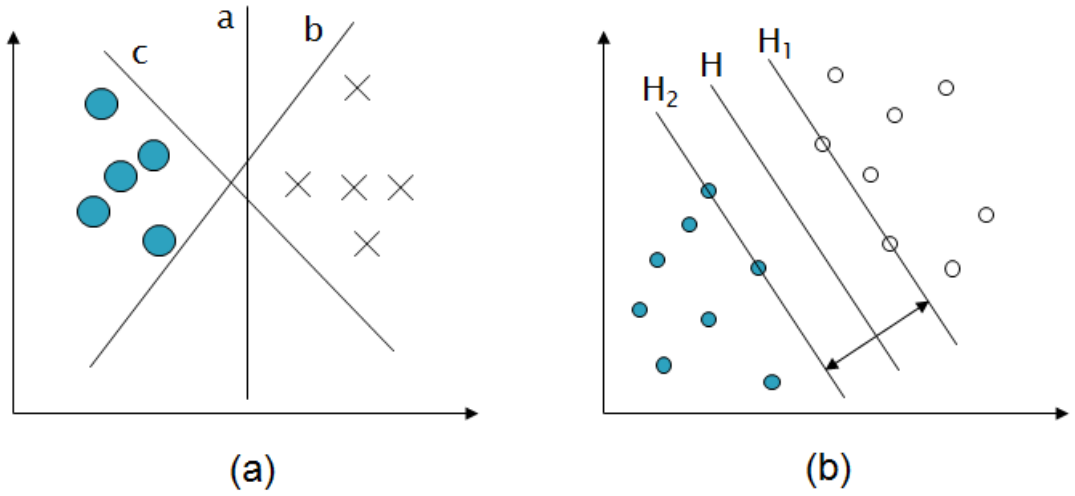


图 2.19: SVM分类示意图

（4）对于待分类向量x，选择某一特定的核函数 $K(x, x_i)$ ，计算：

$$f(x) = sgn(\sum_{i=1}^{n} a_i^* y_i K(x_i, y_i) + b^*) \tag{2.21}$$

式中sgn(*)为符号函数，最终由$f(x)$的取值（+1或-1）决定x属于哪一类。

支持向量机广泛用于风险预测、图像识别以及文本分类等领域。它的优点有：1、较好的泛化能力，能够在数据绛维的同时很好地解决非线性的问题，不会陷于某个局部最优解；2、无需大量的数据训练即可实现小样本的机器学习问题。3、不仅能够用于分类问题，也能用于回归问题。4、支持向量机对训练集之外的数据也具有较好的分类结果，计算复杂度相对较小且结果易于理解。但是，支持向量机具有如下缺点：1、核函数的映射相对较抽象；2、支持向量机对缺失数据较为敏感，当训练集中存在缺失值时可能对分类、预测结果影响较大；3、支持向量机算法结果对调节过程中的参数以及核函数的参数较为敏感，参数不同，结果可能偏差较大。

2.2.4.2 PCA-SVM预报模型

经过前期的数据预处理、预报因子特征工程后，得到的太阳耀斑预报数据集中预报因子的维度为26维。主成分分析（PCA）的“降噪”、“去冗余”的优势在太阳耀斑的快速预报中具有重要作用，它在保留尽可能多的“预报信息”

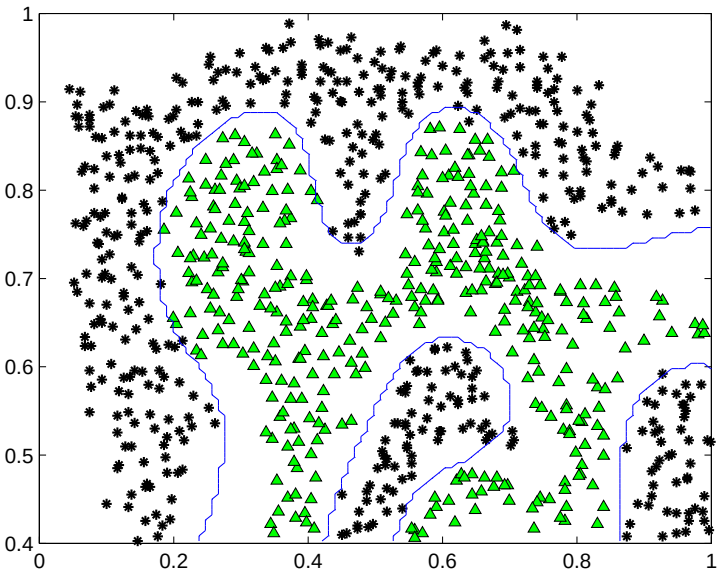


图 2.20: SVM在非线性问题上的应用

的同时大大减少了预报模型的计算量。在各分类算法中，支持向量机（SVM）具有较高的分类能力，核函数的多维映射使得SVM在不加大时间复杂度的同时能够处理太阳耀斑预报这一复杂问题。

[1] 核函数的选择

在SVM中，核函数的引入可以有效避免“维数灾难”，极大地减少分类预测模型中的计算量。常见的核函数有线性核函数、高斯核函数、多项式核函数等。在不同的应用中，往往根据问题的类型、样本数和特征数目的大小以及模型输入特征的不同而选择不同的核函数。表 2.4 展示的是各核函数的主要应用领域。线性核函数主要用于线性可分的情景，它具有参数少、运算快等优势，对于特征数目较大（与样本数目差不多）的数据集，线性核分类器已经具备很好的分类效果。而对于线性不可分的的情景，一般高斯核函数的分类效果优于多项式核函数的分类效果，但是，对于数据变化剧烈的数据集，多项式核反而优于高斯核。在速度上，线性核的速度最快，高斯核次之。多项式核由于参数增加较多，它的速度是三者中最慢的。Sigmoid核函数同样可用于机器学习中的分类预测等问题，但其更适用于事件发生的概率分析。

高斯核作为SVM分类器的核函数具有其特定的优势：（1）高斯核将一个样

表 2.4: 各核函数的应用场景

核函数类型	应用场景
线性核函数	线性分析
高斯核函数	多项式回归
多项式核函数	RBF网络
Sigmoid核函数	概率计算

本映射到更高维度的空间，它不仅适用于线性不可分的情景，同样适用于线性可分的情景。事实上，线性核就是高斯核的一个特例；（2）与多项式核相比，高斯核需要确定的参数要少得多，与高阶多项式核相比，高斯核的时间复杂度要低得多。（3）另外，高斯核具有与Sigmoid核的性能，同样适用于概率的分析。

[2] 模型的训练

在PCA-SVM太阳耀斑预报模型中，数据样本集可分为训练集和测试集，经过前期的一些处理后，太阳耀斑的预报因子共有26 个维度，加上标记是否发生耀斑的标记位，训练集样本共有27个维度。具体的训练、耀斑预报过程如下：

（1）对训练数据进行特征工程处理，得到归一化至[-1, +1]范围的数据集；

（2）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，度量各特征与耀斑发生情况的相关性，计算出数据集的协方差矩阵；

（3）根据计算出的特征向量、协方差矩阵的结果，按照95%信息量的原则选取出前k个主成分构成新基；

（4）根据步骤（3）中的新基构建合适的转换矩阵，然后对原数据进行投影转换，得到转换后的预报因子；

（5）选取高斯核作为本次耀斑预报模型中的核函数，利用支持向量机训练步骤（4）中的转换后的训练集，根据网格测试训练的结果选取最佳模型参数并以此构成全新的耀斑预报模型；

（6）根据步骤（5）中得到的耀斑预报模型对新输入的样本数据进行分类、预报。

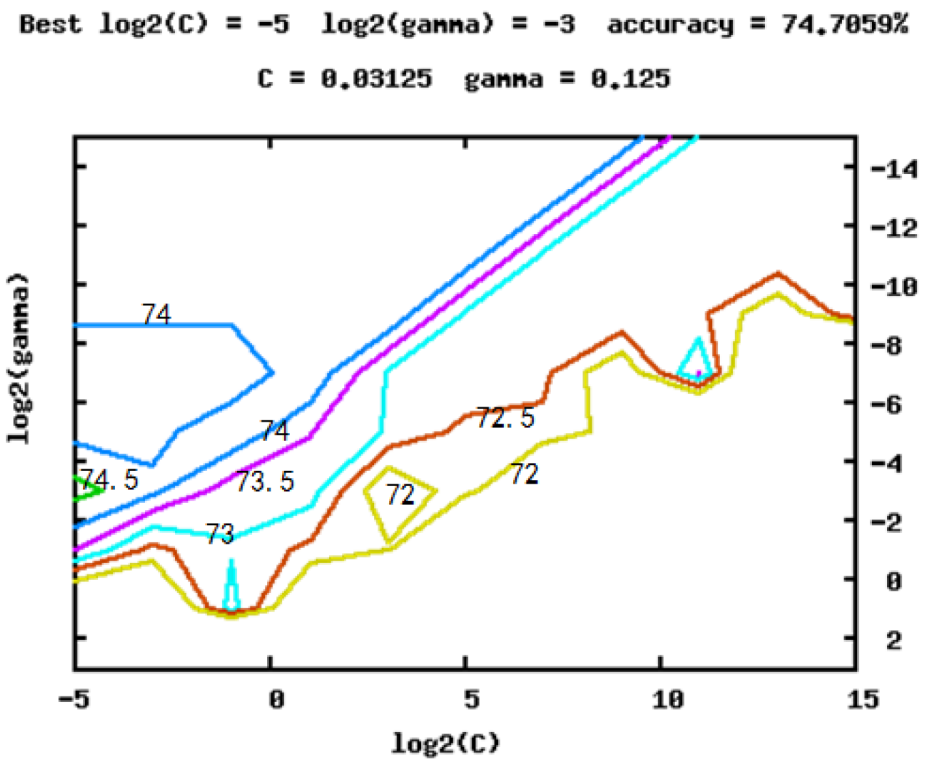


图 2.21: SVM训练进程图

2.3　结果验证

本论文搜集并整理了2014年全年的太阳耀斑发生情况数据。该数据集共有2394个样本数据，其中，未来48小时内发生M 级及以上耀斑的样本数据仅有221个。考虑到正反样本数目的严重失衡，本文随机从反例样本中抽取了等量的正反样本数据并采用十折交叉验证的手段对训练结果进行验证，训练集和验证集的占比分别为80%和20%。

2.3.1　混淆矩阵

机器学习中的二分类问题的分类预测结果可以以混淆矩阵表示（见表 2.5 ）。假设有N个样本，其中，正例样本数为N_p 以及反例样本数为N_n ，则有：

$$N_p + N_n = N \tag{2.22}$$

在预测为正例的样本中，有真实为正例的样本（True Positive, T_P）和真实

表 2.5: 二分类问题的混淆矩阵

	预测正例	预测反例
真实正例	T_p	F_n
真实反例	F_p	T_n

表 2.6: 预报结果对比之混淆矩阵

	文献 [79] 预报模型		PCA-SVM预报模型	
	预测正例	预测反例	预测正例	预测反例
真实正例	125	96	139	82
真实反例	18	203	29	192

为反例的样本（False Positive, F_P）；在预测为反例的样本中，有真实为反例的样本（True Negative, T_N）和真实为正例的样本（False Negative, F_N），它们之间与N_p和N_n的关系有：

$$T_p + F_n = N_p \tag{2.23}$$

$$T_f + F_p = N_f \tag{2.24}$$

它们与N的关系为：

$$T_p + F_n + T_f + F_p = N \tag{2.25}$$

利用同样的预报因子，分别以文献 [79] 中的基于多层感知器的太阳耀斑预报模型进行对比，预报分类的结果如表 2.6 所示。对于442个样本数据，文献 [79] 方法预测为正例的样本数为143 个，其中T_p 个数为125个，预测为反例的样本个数为299个，其中T_n个数为203个；而对于PCA-SVM太阳耀斑预报模型，预测出的168 个正例样本中T_p个数为139，预测出的274个反例样本中T_n个数为192。整体而言，文献 [79] 预报模型预测正确的样本数为328（125+203）个，而PCA-SVM 预报模型预测正确的样本数为331 （139+192）个。

在机器学习领域，精确率（Precision）和召回率（Recall）是两个比较常用的评估标准，精确率被定义为“预测正确的正例数”与“预测为正例的样本

数”的比值，召回率被定义为“预测正确的正例数”在“真正正例的样本数”中的占比。公式表达如下：

$$Pre = \frac{T_p}{T_p + F_p} \tag{2.26}$$

$$Rec = \frac{T_p}{T_p + F_n} \tag{2.27}$$

考虑这两种情况：（1）当将所有样例都预测为正例时召回率能达到100%，但是其精确率只有$\frac{N_p}{N}$；（2）当且仅当一个正例被准确预测时，精确率为100%，但召回率只有 $\frac{1}{N_p}$ 。所以，仅仅从精确率或者召回率来评价一个分类预测模型并不全面，为此，机器学习中引入了F1值的评价参量，F1值被定义为：

$$F1 = \frac{2 * Pre * Rec}{Pre + Rec} \tag{2.28}$$

可计算得，PCA-SVM耀斑预报模型的精确率、召回率和F1值分别为82.7%、62.9%和0.715。相较于文献 [79] 的太阳耀斑预报模型，PCA-SVM预报模型的精确率略低但召回率较高，并且，以F1 这一综合参量来看PCA-SVM预报模型具有更高的效益。对于太阳耀斑的监测系统而言，人们往往希望能够捕捉到每一次的耀斑发生，对于“误捕捉”的成本并不很大。因此，在机器学习领域，PCA-SVM预报模型是一种有效的预报模型。

2.3.2　准确率和虚报率

准确率和虚报率是太阳耀斑预报的评价体系中极其重要的两个评价指标。准确率为“实为正例并被预测为正例的样本数”与“所有预测准确的样本数”的比值（公式 2.29 ），而虚报率为“实为反例却被预测为正例的样本数”在“预测为正例的样本数”中的占比（公式 2.30 ）。表 2.7 显示的是文献 [79] 预报模型和PCA-SVM 预报模型在这两个指标上的对比。

$$Tpr = \frac{T_p}{T_p + T_n} \tag{2.29}$$

$$Fpr = \frac{F_p}{T_p + F_p} \tag{2.30}$$

虽然文献 [79] 预报模型具有相对较低的虚报率，但无论是从预测正例的准确率还是整个样本的准确率（Tpr）来看，PCA-SVM 预报模型都具有较好的预

表 2.7: 各预报模型准确率与虚报率的对比

	文献 [79] 预报模型	PCA-SVM预报模型
准确率(%)	74.2	74.9
虚报率(%)	12.6	17.3

报性能。在太阳耀斑预报这一工程运用中，正例样本的重要性要远高于反例样本，因此，在准确率和虚报率这一评价体系看，PCA-SVM预报模型仍然是一种有效的太阳耀斑预报模型。

2.4 小结

在太阳耀斑的预报中，既有基于太阳物理等专业知识而建立的物理预报方法，又有基于大量观测数据并结合计算机技术而建立出的预报方法。本章节研究的正是这结合观测数据和计算机技术的太阳耀斑预报方法。在这一章节里，首先简要介绍了机器学习方法以及机器学习在具体工程实践中的实现过程，以加深读者对机器学习以及计算机技术的理解。然后，针对太阳耀斑预报这一具体的任务，详细介绍了太阳耀斑的预报因子以及这些预报因子“特征工程”的处理过程：包括数据的预处理以及特征的选择。本文选用的太阳耀斑预报因子主要包括太阳黑子活动参量以及射电流量数据等，对于这些耀斑预报数据，采用了有针对性的数据预处理流程。这一数据预处理过程包括数据无量纲化、针对定量特征的二值化、针对定性特征的编码以及数据缺失值的填充。在介绍机器学习编码技术的同时引入了“One-Hot”编码等技术。

经过特征工程的“特征选择”，本文选取出了一批合适的太阳耀斑预报因子。然后分析了几种分类预报算法并试图建立出一种合适的太阳耀斑预报模型：PCA-SVM太阳耀斑预报模型。在建立这一模型时，主要考量了核函数的选择以及参数的确定，然后根据前期准备好的数据集（含训练集、测试集和验证集）对模型进行训练并最终得到一种太阳耀斑预报模型。

之后，以两种完全不同的评估手段对训练好的模型进行效果评估：包括混淆矩阵评价体系以及准确率和虚报率评价体系。总体上说，本文综合考虑了太阳黑子参量、射电流量等太阳耀斑预报因子，然后在分析各机器学习算法的同时提出了一种新的有效的太阳耀斑预报模型：PCA-SVM太阳耀斑预报模型。

这一耀斑预报模型可用于预测未来48小时内是否会发生M级以上耀斑，可以作为一种太阳活动水平的评估手段。由于特征提取水平的限制，目前这一太阳耀斑预报模型并未考量包括磁场剪切、磁螺度和电流螺度、纵向磁场最大水平梯度、中性线长度和孤立奇点等参量在内的重要预报因子，在今后的研究中可适当添加合适的预报因子，以便进一步提高该模型在太阳耀斑预报中的准确率以及对太阳活动水平评估的能力。

第三章　太阳活动特征检测方法研究

3.1　方法概述

3.1.1　目标检测的概念

与机器学习类似，图像目标检测同样是当前备受关注的前沿方向之一[80][81]。在太阳耀斑的监测中，往往需要预先识别某些特定的太阳活动特征，这就需要利用图像处理等技术来完成对太阳活动特征的检测了。图像目标检测是一种基于几何和统计特征的图像分割[82]、目标识别技术。

图像目标检测任务可以分为目标的分类以及目标的定位这两个关键的子任务。作为高级计算机视觉的一种应用，图像目标检测被应用于许多实际的任务，例如智能化交通、医学监控、军事目标检测等。图像目标检测对于科学研究和工程应用都具有重大意义，随着机器学习方法和计算机技术的不断发展，图像目标检测技术也日趋完善。本章节将利用图像目标检测技术进行太阳各活动特征的检测等应用。

3.1.2　形态学操作

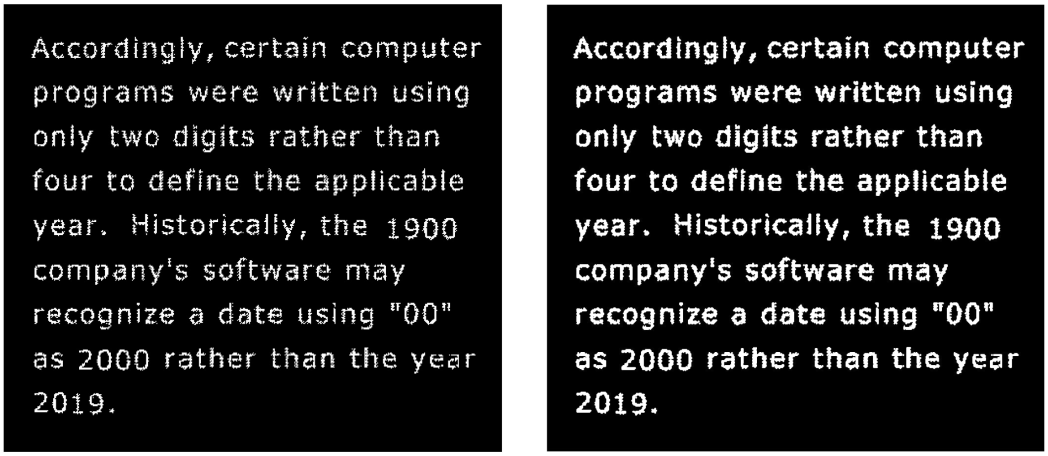
图像形态学操作[83]是在几何形状和结构上提取、度量某些几何结构，达到图像处理、图像识别目的的操作。常见的几种图像形态学操作有膨胀、腐蚀、开运算和闭运算等[84]。

3.1.2.1　膨胀和腐蚀

膨胀和腐蚀是图像形态学中的两种基本操作。膨胀和腐蚀可以寻求图像中的极大值或极小值，同时也可以帮助连接相邻的元素或分割独立的图像元素。通过合适的处理，膨胀和腐蚀可以用于消除某些特定的噪声。

[1] 膨胀

在图像形态学中，膨胀可解释为求局部最大值的操作，是在二值图像中进行“变粗”或者“加长”的操作。形象地说，膨胀的具体实现是由一个结构元素在原图基础上经过一定运算后形成的，图像膨胀后的相当于在原图的基础上



（a） （b）

图 3.1: 形态学操作之膨胀

加上了结构元素后的区域。如图 3.1 所示，图 3.1 （a）为残缺文本的输入图像，经过膨胀处理后文本如图 3.1 （b）所示，相较于膨胀前，文本的高亮区域“加长”了。

[2] 腐蚀

腐蚀与膨胀相反，在图像形态学中对应于求局部最小值的操作。同样地，在一个结构元素的帮助下，腐蚀将侵蚀掉原图的一些精细结构，使原图显得更加平滑。如图 3.2 所示，图 3.2 （a）为具有精细结构的输入图像，经过腐蚀处理后图像如图 3.2 （b）所示，相较于腐蚀前，腐蚀后图像的高亮区域更窄了且丢失了部分精细结构。

3.1.2.2 开运算和闭运算

[1] 开运算

开运算是膨胀腐蚀的组合操作。对于一个图像，先腐蚀后膨胀这一过程被称为开运算。开运算可用于（1）、消除较小的物体；（2）、在图像纤细点处对两个物体进行分离；（3）、在不明显改变物体面积的同时平滑相对较大物体。形态学开运算的数学公式为：

$$A \circ B = (A \ominus B) \oplus B \tag{3.1}$$

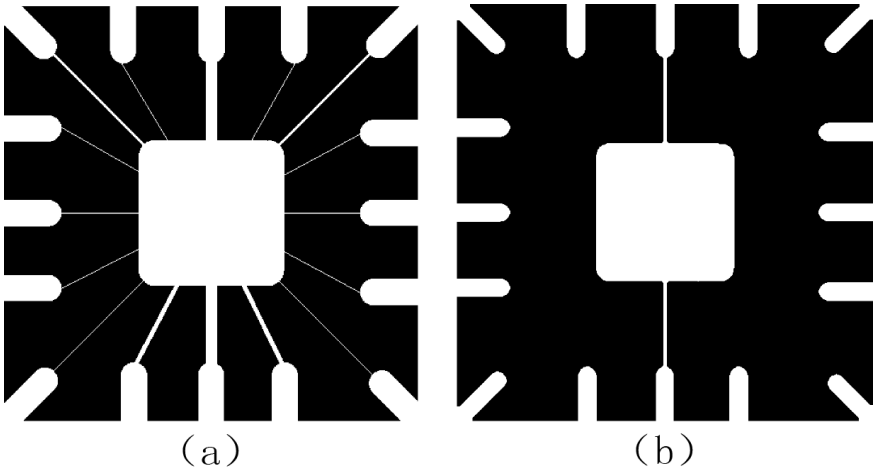


图 3.2: 形态学操作之腐蚀

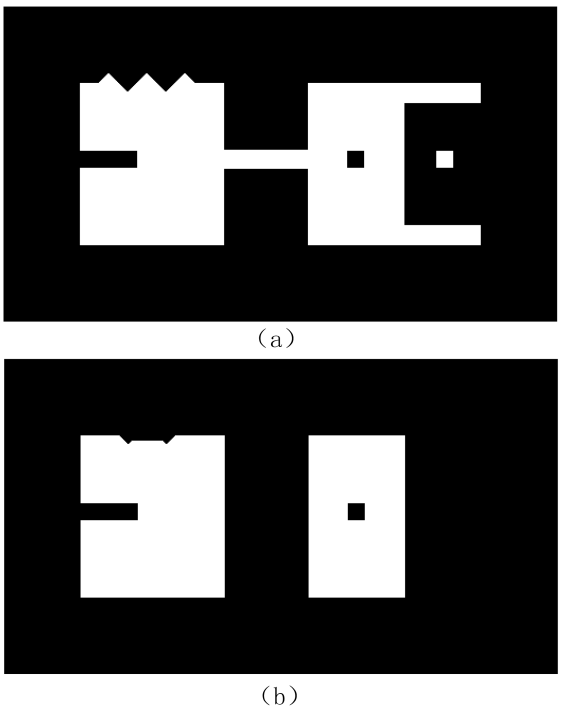


图 3.3: 形态学操作之开运算

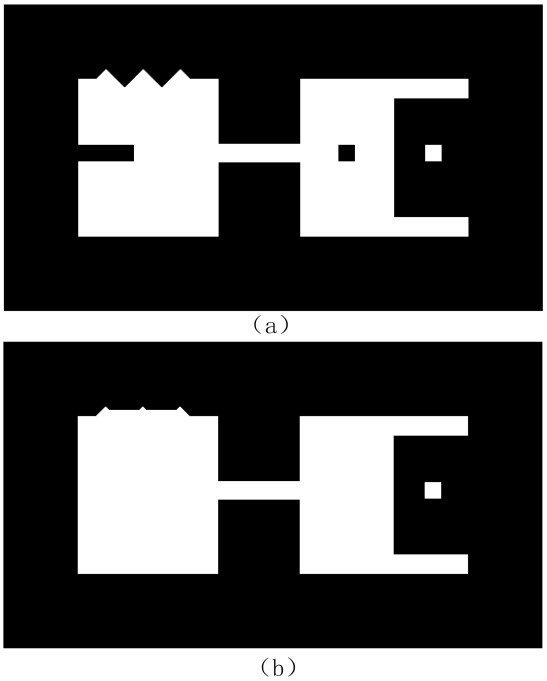


图 3.4: 形态学操作之闭运算

其中，A∘ B 表示A被B的形态学开运算，⊕ 为膨胀操作，⊖ 为腐蚀操作。如图 3.3 所示，图 3.3 （b）是图 3.3 （a）经过开运算处理后得到的结果。

[2] 闭运算

与开运算相反，闭运算即是图像先膨胀后腐蚀的过程。在图像处理中，闭运算的用处有：（1）、用来填充物体内细小空洞，使光亮区域更大；（2）、连接两个相邻的物体；（3）、在不明显改变物体面积的同时平滑其边界。形态学闭运算的公式为：

$$A \bullet B = (A \oplus B) \ominus B \tag{3.2}$$

同样地，A• B 表示A被B的形态学闭运算。如图 3.4 所示，图 3.4 （b）是图 3.4 （a）经过开运算处理后得到的结果。从图中可以看出，闭运算填充了物体内细小的空洞并将狭窄的缺口连接了起来。

3.1.3 边缘特征的操作

在图像目标检测中，边缘是数字图像中强度变化明显的点的集合，图像强

度变化明显通常反映了某些目标的变化或者一些重要的特征，包括目标表面方向上的不连续、深度上的不连续以及目标属性的变化。边缘检测是对于强度间断检测中最为普遍的检测方法，它是图像目标检测，乃至整个计算机视觉中极为重要的一个研究领域。常见的图像边缘检测算法有Roberts算子、Sobel算子和Laplacian算子等。另外，除边缘检测外，图像边缘的保持（如双边滤波）同样是数字图像增强处理过程中重要的一环。

[1] Roberts算子

Roberts算子使用公式 3.3 所示的A和B两个局部差分掩膜寻找边缘，图像中各点利用这两个掩膜做卷积，如式 3.4 所示：

$$A=\begin{bmatrix}-1 & 0\\ 0 & 1\end{bmatrix}, B=\begin{bmatrix}0 & -1\\ 1 & 0\end{bmatrix} \tag{3.3}$$

$$g(x,y)=[\sqrt{f(x+1,y+1)}-\sqrt{f(x,y)}]^2+[\sqrt{f(x,y+1)}-\sqrt{f(x+1,y)}]^2 \tag{3.4}$$

其中f(x,y)、f(x,y+1)、f(x+1,y)和f(x+1,y+1)表示图像中的4邻域，g(x,y)为经Roberts算子更新后的像素值。Roberts算子对于边缘定位具有较高的精度，但由于没有经过图像平滑处理，不能抑制噪声。对于如图 3.9 所示的太阳图像，Roberts算子的边缘检测效果如图 3.5 所示。从图中可以看出，Roberts算子较好地检测出了太阳日面边缘，对于谱斑、暗条等太阳活动特征的边缘也具有一定效果，但是，Roberts对噪声的抑制能力较差。

[2] Sobel算子

作为一种一阶微分算子，Sobel算子使用公式 3.5 所示的A和B两个卷积核计算水平边缘以及垂直边缘，然后以这两个卷积的最大值作为该点的输出。

$$A=\begin{bmatrix}-1 & -2 & -1\\ 0 & 0 & 0\\ 1 & 2 & 1\end{bmatrix}, B=\begin{bmatrix}-1 & 0 & 1\\ -2 & 0 & 2\\ -1 & 0 & 1\end{bmatrix} \tag{3.5}$$

假设图像的8邻域如式 3.6 所示，则Sobel算子的水平边缘响应和垂直边缘响应分别如式 3.7 、 3.8 所示。

$$N_8=\begin{bmatrix}z_1 & z_2 & z_3\\ z_4 & z_5 & z_6\\ z_7 & z_8 & z_9\end{bmatrix} \tag{3.6}$$

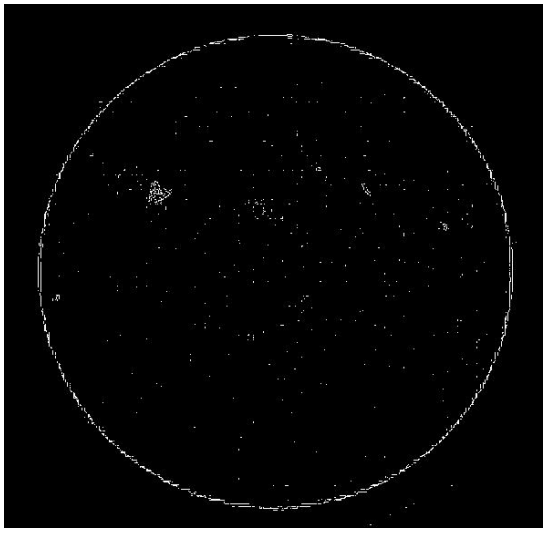


图 3.5: Roberts算子的边缘检测效果

$$G_x = (z_7 + 2 * z_8 + z_9) - (z_1 + 2 * z_2 + z_3) \tag{3.7}$$

$$G_y = (z_3 + 2 * z_6 + z_9) - (z_1 + 2 * z_4 + z_7) \tag{3.8}$$

式中 G_x 和 G_y 分别为水平和垂直方向上的响应。Sobel算子的最终响应g为：

$$g = (G_x^2 + G_y^2)^{\frac{1}{2}} \tag{3.9}$$

Sobel算子隐含有加权平均的操作，在边缘检测过程中有一定的噪声抑制作用。但是，由于Sobel算子只采用了水平垂直两个方向的模板，只考虑了水平垂直这两个方向上的响应，因此，Sobel算子在处理纹理较为复杂的图像时效果并不是很理想。同样是对于图 3.9 的太阳色球层图像，Sobel的检测效果如图 3.6 所示。Sobel算子对于电子噪声等具有很好的抑制效果，且对于太阳日面边缘，Sobel算子能较好地检测其边缘，但对于谱斑等具有较为复杂纹理的太阳活动特征的检测效果并不很理想。

[3] Laplacian算子

Laplacian算子是一种二阶微分算子，它一般不以其原始形态用于边缘检测。LoG（Laplacian of a Gaussian）边沿检测器是最常见的Laplacian 算子检测器之一。考虑高斯函数：

$$h(r) = -e^{-\frac{r^2}{2\sigma^2}} \tag{3.10}$$

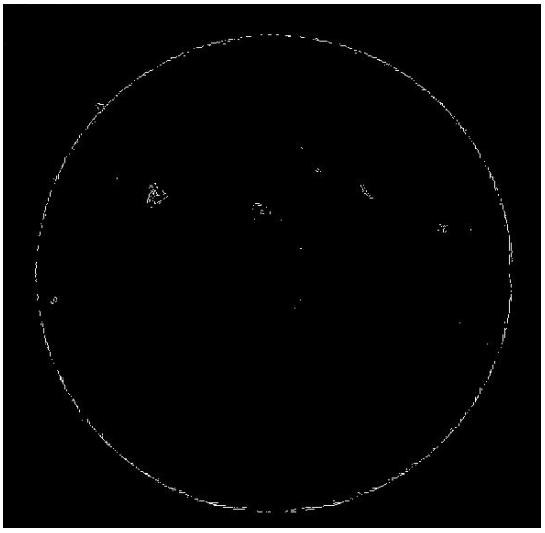


图 3.6: Sobel算子的边缘检测效果

高斯函数与一幅图像做卷积会使图像变得模糊，图像的模糊程度由标准偏差σ决定。该函数的Laplacian算子为：

$$\nabla^2 h(r) = -[\frac{r^2 - \sigma^2}{\sigma^4}]e^{-\frac{r^2}{2\sigma^2}} \tag{3.11}$$

上述函数即LoG边缘检测器中的LoG函数，该函数对图像进行卷积滤波使图像变得平滑的同时也一定程度上减少了噪声，另外，由于Laplacian 算子的性质，该算子定位边缘就是找到两个边缘间的零交叉。Laplacian算子不具有方向性，对图片强度突变敏感，定位精度较高，但同时对噪声也较为敏感。对于如图 3.9 所示的太阳图像，LoG边缘检测器的检测效果如图 3.7 所示。从图中可以看出，LoG边缘检测器较好地检测出了太阳日面以及其他太阳活动特征的边缘，但是，对于噪声的抑制能力不及Sobel算子。

[4] 双边滤波

双边滤波（Bilateral filter）是一种既考虑了像素位置、强度的差异，又具有边缘保持特性的非线性滤波器。顾名思义，双边滤波由两个函数组成，即在高斯滤波器（Gaussian Filter，图 3.8）的基础上增加了一个与像素强度相关的函数。高斯滤波器的权重只与像素的空间距离有关，表达式如公式 3.12 所示：

$$W_{ij} = \frac{1}{K} exp(-\frac{|x_i - x_j|^2}{\sigma^2}) \tag{3.12}$$

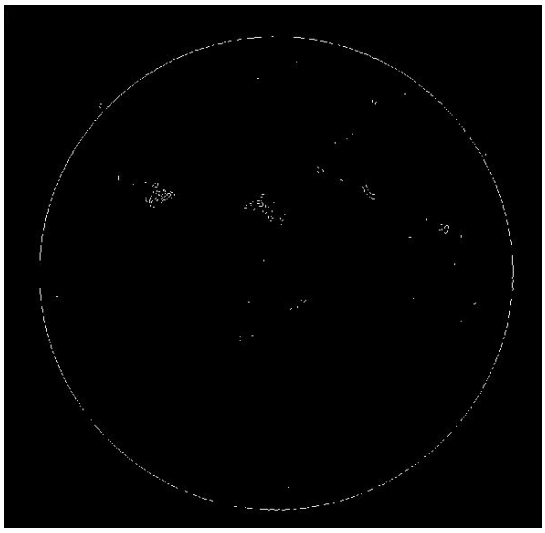


图 3.7: Laplacian算子的边缘检测效果

其中，i和j是图像的像素索引，W_{ij}是像素i与j之间的权重，K是归一化的常量。从公式中可以看出，高斯滤波器的权重只与距离有关，离i点相同的任意j点对i点有着相同的滤波效果。高斯滤波对于高频细节没有很好的保持效果，但双边滤波与此有着显著区别，由于增加了一个与像素强度相关的函数，双边滤波不仅有着“图像平滑”的作用，而且具有一定的“边缘保持”特性，其滤波表达式如下：

$$W_{ij} = \frac{1}{K} exp(-\frac{|x_i - x_j|^2}{\sigma^2}) exp(-\frac{|I_i - I_j|^2}{\sigma_r^2}) \tag{3.13}$$

同样地，i和j是图像的像素索引，I_i和I_j分别代表图像在i、j点的像素强度。由表达式可知，在强度差距较大的地方（高频细节区域，或称“边缘”），滤波器的权重会减小，滤波效应也将变小。总地来说，双边滤波在像素强度变换不大的区域有类似于高斯滤波的平滑效果，而在图像高频细节区域则有着姣好的边缘保持效果。

虽然双边滤波有着高斯平滑以及“边缘保持”等优点，但同时也存在着一些不足。若某个像素点周围有很少强度相似的像素时，滤波器的高斯加权平均值并不稳定，有可能造成梯度反转的现象。另一方面，双边滤波的计算效率较低，其时间复杂度为$O(Nr^2)$，当窗口半径r 较大时将大大增加滤波器的计算时间。

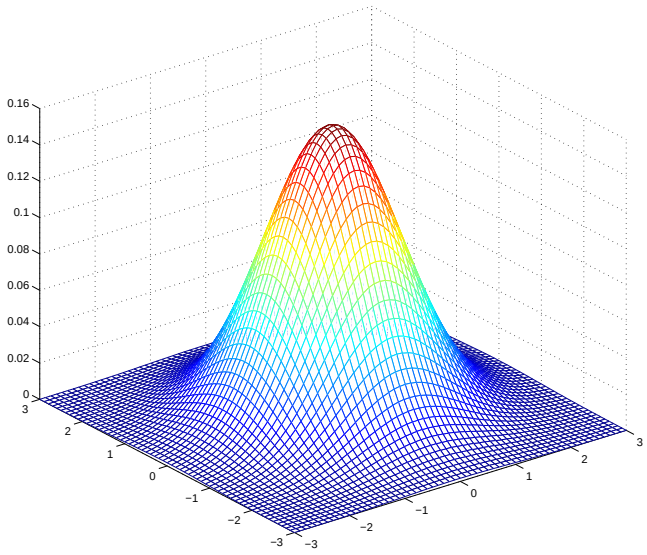


图 3.8: Gaussian滤波函数

3.1.4　阈值处理

图像阈值处理可以从复杂图像中获取结构更加简单、像素更加单一的图像。阈值处理方法中的阈值可分为全局性的阈值和在某个局部区域块使用的局部阈值。根据阈值来源的不同，全局阈值或局部阈值又可分为简单阈值、自适应阈值等。

3.1.4.1　简单阈值

简单阈值即直接选取某一个全局变量作为图像阈值处理中的阈值。在简单阈值处理中，可将整幅图像分成了一幅只有0和1 存在的二值图像。简单阈值处理又可分为简单二值化处理和二值反转处理。例如，对于式 3.14 中的图像A，当选定的简单阈值为0.5时，简单二值化处理后的图像如式子B所示。

$$A = \begin{bmatrix} 0.1 & 0.2 & 0.3 \\ 0.3 & 0.4 & 0.4 \\ 0.6 & 0.7 & 0.8 \end{bmatrix}, B = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 1 & 1 & 1 \end{bmatrix} \tag{3.14}$$

而二值反转处理后的图像如式中 3.15 C 所示，可以看出，图像B与C是完

全相反的。

$$A = \begin{bmatrix} 0.1 & 0.2 & 0.3 \\ 0.3 & 0.4 & 0.4 \\ 0.6 & 0.7 & 0.8 \end{bmatrix}, C = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 0 & 0 & 0 \end{bmatrix} \tag{3.15}$$

3.1.4.2 自适应阈值

图像与图像之间能够差距很大，简单的阈值往往不能很好地分割图像。在实际工程应用中，人们往往需要根据图像的不同而设置不同的阈值，即自适应阈值。OTSU算法是一种较为“完美”的自适应阈值算法，它又被称为大津算法，是由日本学者OTSU 于20 世纪70 年代提出的一种高效的图像二值化算法 [85] [86] 。严格意义上说，OTSU 算法也是一种自适应阈值算法。它将图像分成前景和背景两个部分，当取最佳阈值时，前景和背景两个部分之间的灰度值差异最大且各部分内的灰度值差异最小，灰度值差异的大小即方差的大小。所以，通过OTSU 算法可以自动选取出一个“自适应阈值”对图像二值化。现有输入图像$I(x, y)$ 以及将图像划分为前景和背景的阈值T，假如前景和背景的像素个数分别为N_1 和N_2，图像I 的大小为$M \times N$，可知：

$$N_1 + N_2 = M \times N \tag{3.16}$$

若w_1和w_2分别表示前景和背景像素占整个图像的比例，则有以下关系：

$$w_1 = \frac{N_1}{M \times N} \tag{3.17}$$

$$w_2 = \frac{N_2}{M \times N} \tag{3.18}$$

$$w_1 + w_2 = 1 \tag{3.19}$$

假设前景和背景的平均灰度值分别为μ_1 和μ_2，则整幅图像的平均灰度值和方差分别为：

$$\mu = w_1 * \mu_1 + w_2 * \mu_2 \tag{3.20}$$

$$g = w_1 * (\mu_1 - \mu)^2 + w_2 * (\mu_2 - \mu)^2 \tag{3.21}$$

将公式3.20 和公式3.21代入后可得，整幅图像方差g为：

$$g = w_1 * w_2 * (\mu_2 - \mu_1)^2 \tag{3.22}$$

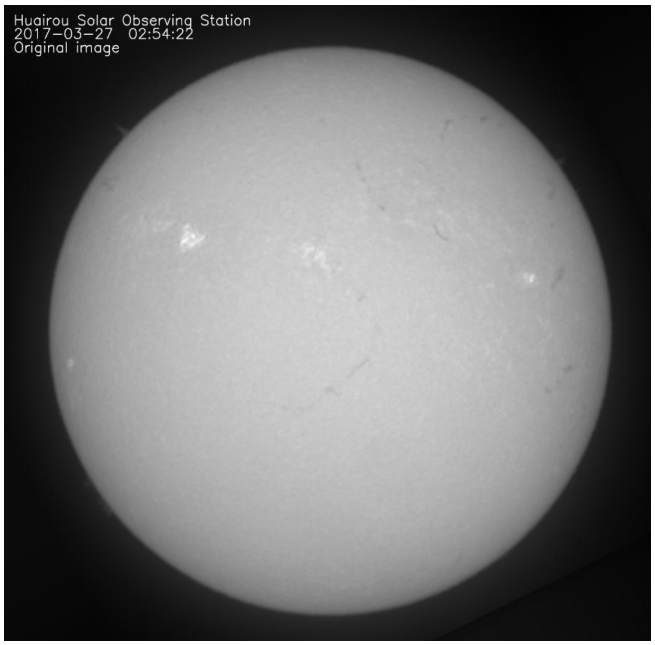


图 3.9: 全日面色球层图像

遍历所有可能的阈值T，使得方差g最大的阈值即为OTSU阈值。OTSU算法有着算法简单，不受图像的亮度以及对比度等因素影响的优点，广泛被认为是图像分割中阈值选取的理想“自适应阈值”算法。

3.2　数据处理与方案设计

与常规目标检测系统不同，太阳活动特征的检测更倾向于关注某些特定太阳活动特征的具体结构与位置，而较少地关注具体活动特征的具体分类（如太阳黑子、活动区、谱斑和耀斑等）。太阳活动特征的检测可以分为活动特征的分类以及活动特征的定位。在太阳活动特征检测中，除了需要区分出具体的活动特征外，还需要定位到这些具体特征的大小、位置和面积等参量。

3.2.1　数据介绍

中国科学院国家天文台怀柔太阳观测基地（Huairou Solar Observing Station，HSOS）目前配有多台套先进的太阳观测设备。利用这些观测设备，可以获取每日的色球层图像（图 3.9 ）和光球层图像（图 3.10 ）。全日面$H\alpha$图像来源于HSOS 的网站 http://sun.bao.ac.cn/hsos_datas/full_disk/h-alpha/ ，另外，可以

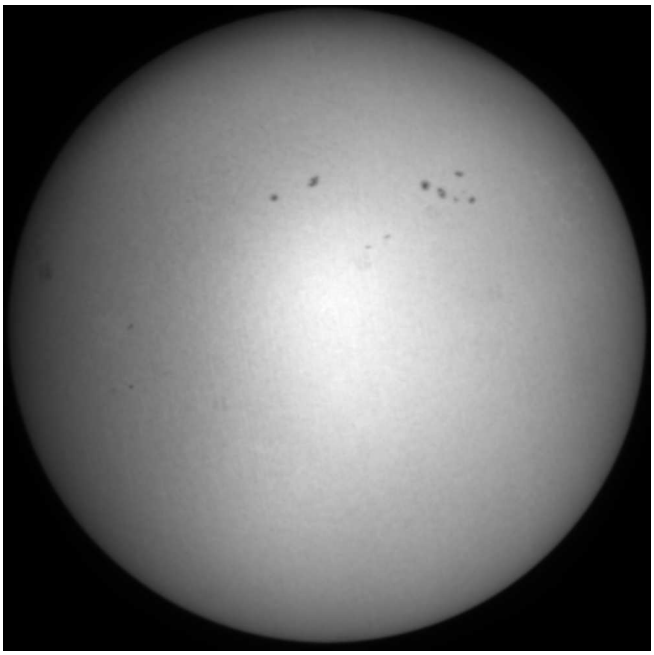


图 3.10: 全日面光球层图像

在 http://sun.bao.ac.cn/hsos_datas/full_disk/magnetogram/ 处获取全日面太阳磁场数据。

3.2.2 图像预处理

数字图像的噪声主要有：（1）由电子元器件引起的高斯噪声；（2）图像切割、图像变换时引起的椒盐噪声；（3）图像在信道传输的过程中引起的加性噪声等。对于这些噪声，可以采用均值滤波、中值滤波或者高斯滤波等滤波器去除。考虑到获取的太阳图像的性质，这里采用均值滤波即可。

3.2.3 图像特征的提取

在太阳图像中，人们关注的目标主要有太阳黑子、活动区、谱斑和耀斑等。不同目标的检测方式各不相同，在图像目标检测技术中，图像特征与图像目标息息相关。一个良好的图像特征应该尽可能地表现出目标的本质特性，能最大限度地独立于目标检测时的环境条件。在目标检测中，不同的特征提取方式可生成不同的图像特征，考虑到太阳活动特征检测的独特性质，本节将以引导滤波来开展太阳图像的特征提取、处理过程。

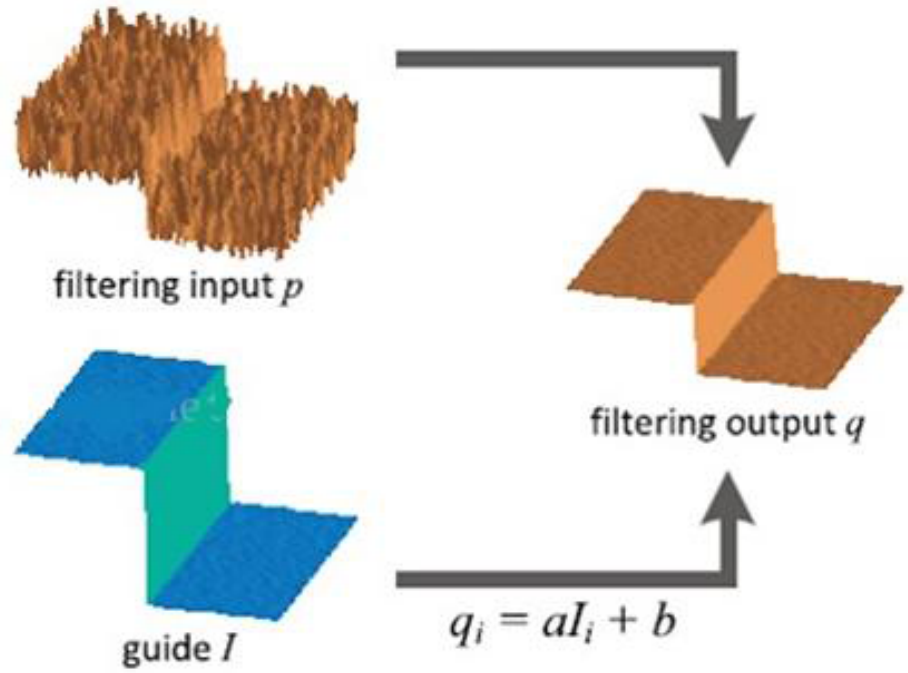


图 3.11: 引导滤波示意图

引导滤波 [87] [88] （Guided Filter）是由何凯明等人于2010年提出的一种边缘保持或结构保持滤波器。相对于双边滤波，引导滤波有着更低的时间复杂度O(N) 和更好的边缘保持特性且不会有梯度反转的风险。在特定引导图像或者自身的引导下，引导滤波可用于平滑滤波、去雾或边缘增强等。

引导滤波的滤波示意图如图 3.11 所示，对于输入图像p，有引导图像I，输出图像q。作为一种结构保持或边缘保持滤波器，引导滤波的理论基础是局部线性模型，该模型认为，在以k为中心的掩膜窗w_k内，输出图像q 是输入图像p的线性转换。输出图像与输入图像的关系如下：

$$q_i = a_k * I_i + b_k, \ \forall i \in w_k \tag{3.23}$$

其中，a_k 和b_k 是在窗口w_k下的常量。引导滤波的求解过程是在该窗口下找出合适的参数最小化输入图像p 和输出图像q 之间的差异。该差异函数$E(a_k, b_k)$被定义为：

$$E(a_k, b_k) = \sum_{i \in w_k}((a_k * I_i + b_k - p_i)^2 + \epsilon * a_k^2) \tag{3.24}$$

为了防止参数a_k过大，式中添加了正则项参数ϵ。根据一系列数学计算可得：

$$a_k = \frac{\frac{1}{|w|}\sum_{i\in w_k} I_i * p_i - \mu_k * \bar{p}_k}{\sigma_k^2 + \epsilon} \tag{3.25}$$

$$b_k = \bar{p}_k - a_k * \mu_k \tag{3.26}$$

其中σ_k^2 和μ_k分别为引导图像I在窗口w_k内的方差和均值，|w|为窗口w_k 内像素的总数目。最终，输出图像q可表示为：

$$q_i = \frac{1}{|w|}\sum_{k:i\in w_k}(a_k * I_i + b_k) = \bar{a}_i * I_i + \bar{b}_i \tag{3.27}$$

$\bar{a}_i$ 和$\bar{b}_i$ 分别为在窗口w_k内a_k和b_k的均值。

当使引导图像I等于输入图像p时，引导滤波可作为边缘保持滤波器。分别考虑这两种情况：（1）、在像素强度变化很小的区域，有a 近似于0而b近似于p，此时引导滤波器相当于加权均值滤波，可以平滑部分噪声；（2）、而在高细节的图像区域，有a 近似于1而b 近似于0，此时滤波器输入输出前后差异很小，在这一区域引导滤波有助于保持边缘。

与双边滤波相比，引导滤波在滤波效果上与双边滤波效果差不多。但是在一些细节上，引导滤波的滤波效果较好。引导滤波的时间复杂度是与窗口大小无关的，因此若使用大窗口处理输入图像时，引导滤波具有更高的效率。另外，由于对亮度渐变不敏感，对突变敏感，引导滤波对临边昏暗现象（solar limb darkening）也具有较好的抑制作用。

3.2.4 太阳活动特征的检测

3.2.4.1 日面拟合

HSOS获取的fits文件中太阳日面的大小位置并不固定，为了确定太阳的准确信息，有必要预先确定太阳的中心和半径，即本节中的日面拟合。由于太阳日面为一明亮圆形，日面与背景的灰度值具有较明显差异，为此可以考虑采用阈值处理办法将背景与日面分割。具体的日面拟合过程如下。

日面拟合的过程主要包括图像预处理、日面提取、日面位置和半径确定等过程。以太阳色球层图像为例，具体的日面拟合过程如下：（1）图像预处理：由于电子噪声、天气等原因，获取到的太阳图像并不总是一个干净的日面，为

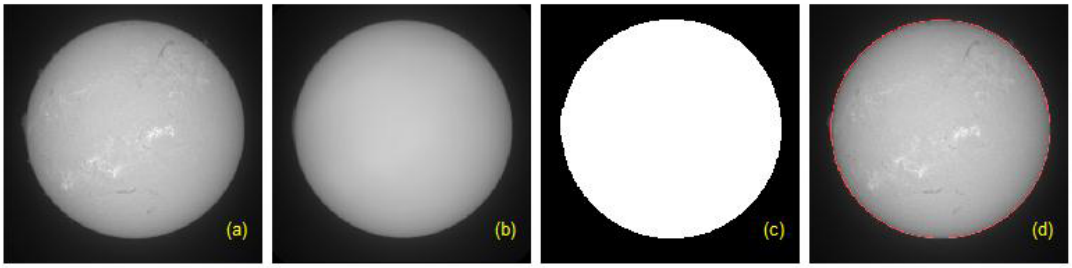


图 3.12: 色球层日面拟合

了便于进一步的日面的处理，首先，利用一个中值滤波对原始图像 3.12 （a）进行去噪处理。考虑到怀柔观测基地的日面半径约为1140像素且每帧图像像素为2712×2712，此处采用的中值滤波窗口大小为320×320。经过去噪处理后得到的干净日面图像如图 3.12 （b）所示。

（2）日面提取：如图 3.13 所示，在太阳图像直方图中，太阳日面具有较高灰度值而背景具有较小的灰度值，因此，可以根据他们灰度值的差距，利用图像二值化将太阳日面与背景分离开，即将大于某一个阈值的像素的设为1而小于某个阈值的像素点设为0。经过对一系列图像的测试，测试发现将阈值设为0.2倍的I_{avg}是比较合适的（其中，I_{avg}为整个太阳图像的平均灰度值）。最后，获得的二值图像如图 3.12 （b）所示。

（3）日面位置和半径的确定：经过霍夫变换可以定位到上述二值图像中的圆形，即太阳日面；然后根据计算行与列的和，最终可以得到太阳日面的中心以及太阳日面的半径。根据已确定的日面中心和半径，可以在色球层图像中定位太阳日面如图 3.12 （c）。最终确定的太阳日面结果如图 3.12 （d）所示。

3.2.4.2　太阳谱斑的检测

确定太阳图像的日面中心与半径后，可以对太阳活动特征进行检测了。太阳黑子的检测过程包括图像预处理、太阳活动特征增强、阈值处理和谱斑的确定，整个过程如图 3.14 所示。具体检测步骤如下：

（1）图像预处理：由于电子设备、天气状况、温度、气候等原因的影响，望远镜观测到的太阳图像的平均亮度并不相同。为了下一步的图像处理，需要在预处理阶段将太阳图像归至于一个合适的亮度范围内，所以，首先需要利用最小最大正则化将太阳图像归一化。通过大量数据测试，发现太阳图像灰度值

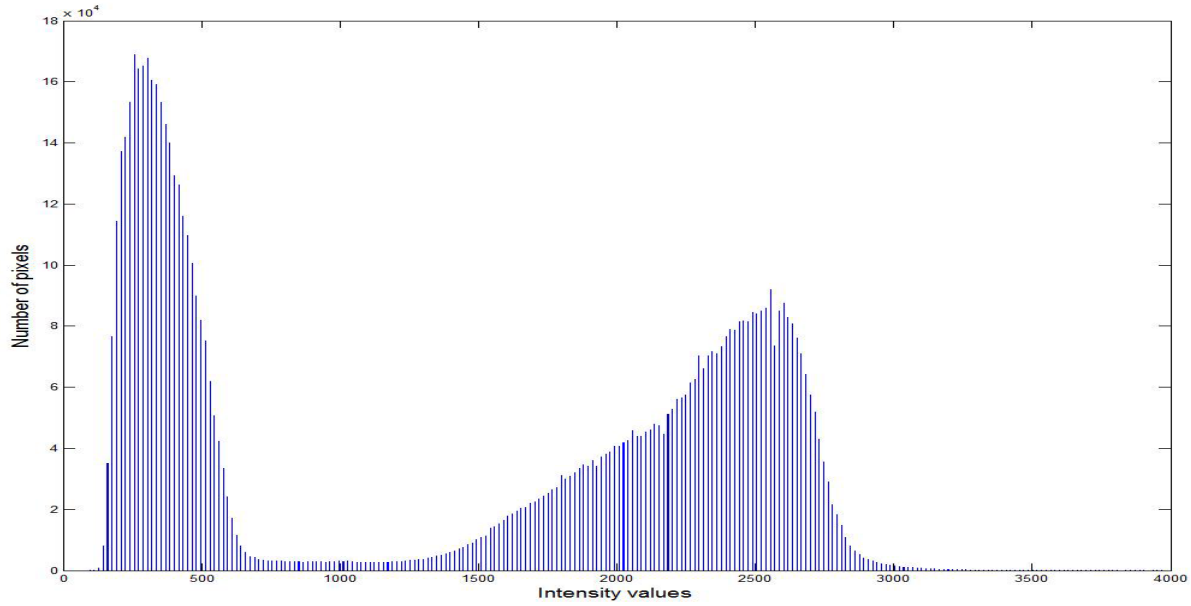


图 3.13: 太阳图像的直方图

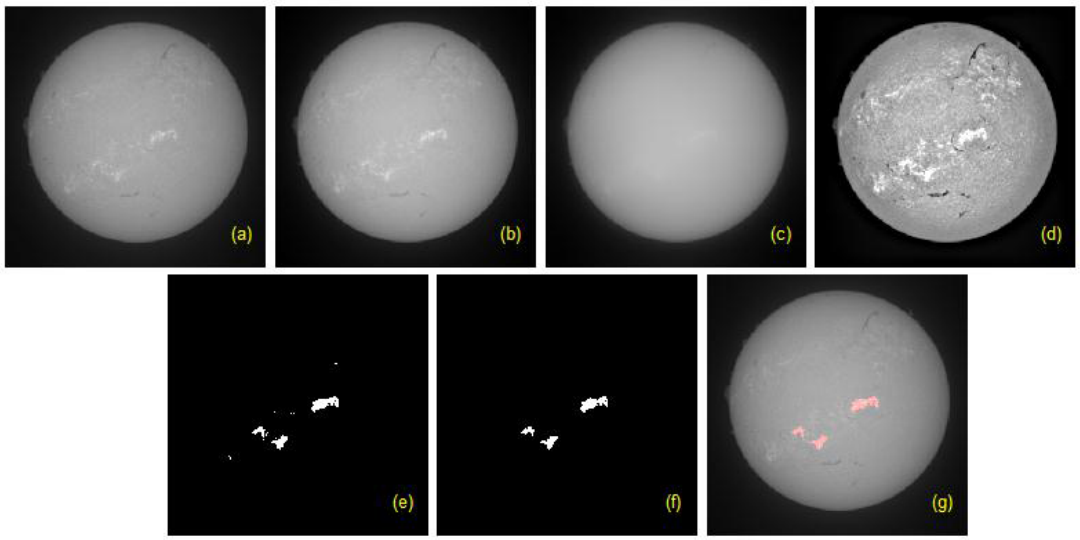


图 3.14: 太阳谱斑的检测

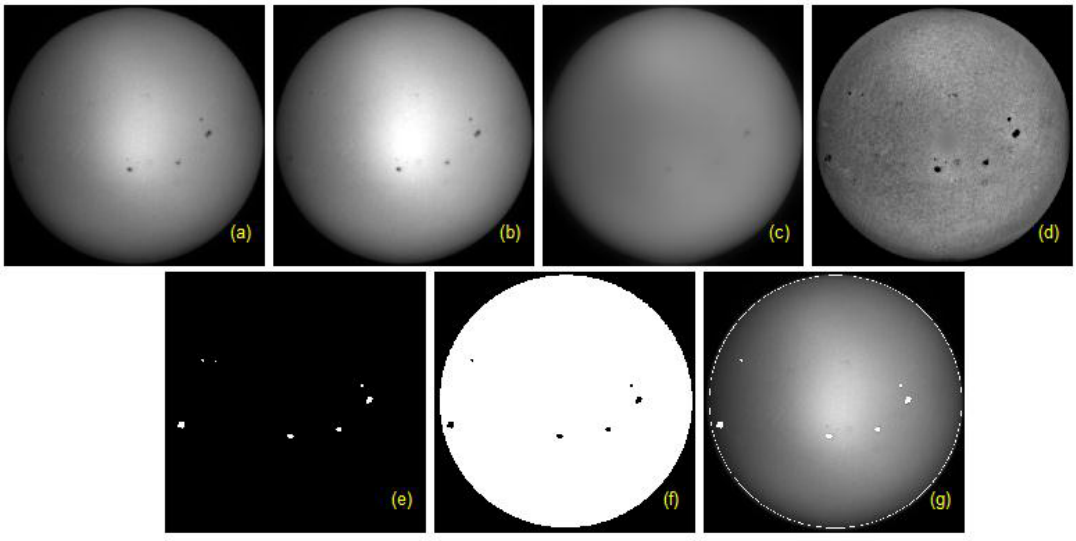


图 3.15: 太阳黑子的检测

归一化至0.42 是比较合适的。另外，加上一个15×15的中值滤波可以减少噪声对图像质量的影响。图像预处理后的结果如图 3.14 （b）所示。

（2）太阳活动特征增强：为了减少背景噪声以及太阳临边昏暗效应的影响，本文引入了一种良好的结构保持滤波器——引导图像滤波器。

（3）阈值处理：阈值处理包括图像二值化以及根据分割后的区域得到候选的谱斑区域块。Otsu算法是一种健壮的自适应图像分割算法，它可以在无预先设定阈值的情况下将图像分割为背景和检测目标两个部分。考虑到本研究需要的是太阳日面上的几个谱斑区域而不是整个日面，将阈值设定为Otsu自适应阈值的2.5倍。图像二值化的结果如图 3.14 （e）所示。分割出来的这几个区域块则是候选谱斑区域。

（4）谱斑的确定：在前面的过程中得到的是候选谱斑区域集，去除噪声等区域后，可以得到最终的谱斑区域，结果如图 3.14 （f）所示。将得到的谱斑区域至于原始图像后，得到的结果如图 3.14 （f）。

3.2.4.3　太阳黑子的检测

太阳黑子的监测与太阳谱斑检测极其类似。以7×7的中值滤波预处理后，可以得到一个较为干净的太阳日面；然后利用引导滤波对其进行图像增强，根据Otsu算法得到候选区域，最后根据这些区域以及阈值后，得到最终的检测结

表 3.1: 太阳谱斑自动检测结果

Day	Number of plages (manual method)	Number of plages (automatic method)	MNA	ANM
2013-12-1	8	8	0	0
2013-12-2	7	5	2	0
2013-12-3	8	6	2	0
2013-12-5	9	7	2	0
2013-12-6	6	6	0	0
2013-12-7	7	6	1	0
2013-12-9	6	5	1	0
2013-12-10	8	7	1	0
2013-12-11	9	8	2	1
2013-12-12	8	7	1	0
2013-12-13	7	6	1	0
2013-12-14	7	5	2	0
2013-12-15	6	5	1	0
2013-12-16	5	5	0	0
2013-12-17	6	5	1	0
2013-12-18	6	4	2	0
2013-12-19	6	4	2	0
2013-12-20	6	4	2	0
2013-12-21	4	3	1	0
2013-12-22	5	5	0	0
2013-12-23	4	3	1	0
2013-12-26	4	4	0	0
2013-12-28	4	3	1	0
2013-12-29	4	2	2	0
2013-12-31	4	5	0	1
TOTAL	154	128	28	2

果如图 3.15 所示。

3.3 结果验证

为了检测太阳活动特征的有效性和准确率，本文将自动检测结果与人工检测结果进行了对比。以HSOS的2013年12月共25天的测试数据为例，太阳谱斑检测和太阳黑子检测的对比结果分别如表 3.1 和表 3.2 所示。表中第一列为太阳图像的观测日期，第二第三列分别为人工方式和程序自动检测到的太阳活动特征的数目（谱斑的数目或者太阳黑子的数目），第四列为人工方法检测到但程序未检测到的个数（Manual Not Automatic，MNA），最后一列为程序检测到但人工方法未检测到的个数（Automatic Not Manual，ANM）。

表 3.2: 太阳黑子自动检测结果

Day	Number of sunspots (manual method)	Number of sunspots (automatic method)	MNA	ANM
2013-12-1	12	11	1	0
2013-12-2	13	12	1	0
2013-12-3	10	10	0	0
2013-12-5	7	7	0	0
2013-12-6	7	6	1	0
2013-12-7	5	5	0	0
2013-12-9	5	4	1	0
2013-12-10	10	8	2	0
2013-12-11	11	11	0	0
2013-12-12	8	8	0	0
2013-12-13	10	9	1	0
2013-12-14	9	9	0	0
2013-12-15	7	7	1	1
2013-12-16	8	8	0	0
2013-12-17	10	10	0	0
2013-12-18	9	10	0	1
2013-12-19	10	9	1	0
2013-12-20	11	11	0	0
2013-12-21	10	9	1	0
2013-12-22	8	8	0	0
2013-12-23	6	5	1	0
2013-12-26	6	6	0	0
2013-12-28	6	6	0	0
2013-12-29	8	6	2	0
2013-12-31	8	7	1	0
TOTAL	214	202	14	2

若以人工检测数目为基准，则检测准确率R_{pre}与错误率R_{err}可分别被定义为：

$$R_{pre} = \frac{N_{auto} - N_{ANM}}{N_{manu}} \tag{3.28}$$

$$R_{err} = \frac{N_{ANM}}{N_{manu}} \tag{3.29}$$

其中N_{auto}和N_{manu}分别为程序自动检测到和人工检测到的数目，N_{ANM}为程序自动检测到但人工未检测到的数目。经计算得谱斑检测准确率和错误率分别为81.8%和1.3%，太阳黑子则分别为93.5%和0.9%。从实验结果可以看出，该太阳活动特征检测程序有较高的准确率以及相对较低的错误率。

本文分析了自动检测程序产生错误的原因，对于人工未检测到但程序检测到的样例，主要原因是太阳图像噪声的干扰以及特征提取过程中的误检测；而对于人工检测到但程序未检测的的样例，本课题认为产生的原因有以下几

点：（1）噪声的干扰使得程序未能检测到；（2）在判断个数时，人工方法将其分为两个甚至多个但程序只将其归为了一个；（3）程序的图像处理过程中丢失部分信息。综合检测准确率和错误率来看，该检测程序仍然是一个有效的太阳活动特征检测方法。

3.4 小结

在这一章节，本文首先引入了图像目标检测的概念，接着针对HSOS的太阳活动特征的检测，逐步介绍了太阳活动特征检测过程中的图像处理（尤其是特征提取）的过程以及相关的算法技术等。最后，根据前期的技术成果，提出了一种有效的太阳活动特征检测办法并将这种检测办法与人工检测结果进行了对比。相比于人工检测结果，该自动检测准确率具有良好的检测效果（检测准确率大于80%，测错误率低于1%）。上一章节基于文本的太阳耀斑预报可以作为太阳活动水平的评估手段，本章节基于图像的太阳活动特征检测可应用于耀斑的捕捉定位，“基于文本的预报”和“基于图像的检测”可共同服务于下一章节的“太阳耀斑实时智能化监测系统的设计与实现”，具有一定的互补优势。

第四章　太阳耀斑实时智能化监测系统的设计与实现

在太阳耀斑发生时，耀斑的亮度在数十秒甚至几秒的时间内快速地增加几倍甚至几十倍，同时其面积也可能迅速增大。面对这种迅速的、急剧的变化，常规观测设备和方法在进行耀斑的观测时常常显得力不从心，很容易产生数据溢出或者错过耀斑数据的情况，因此，在太阳耀斑的实时监测过程中，自动判别太阳耀斑的发生并由此实时修改太阳观测的观测模式变得尤为必要 [89] 。

4.1　系统总体设计

目前怀柔太阳观测基地拥有35CM、60CM局部磁场和全日面磁场望远镜等多台套太阳望远镜，为更好地服务于太阳耀斑的实时监测与高分辨观测，本文在怀柔太阳实时观测系统以及本文前期工作的基础上，完成了太阳耀斑实时监测系统的算法开发以及功能结构的设计。该太阳耀斑观测系统能实时监测太阳的活动情况并根据实时数据提取出合适的特征，然后将这些特征参量代入训练好的太阳耀斑预报模型，以此评估太阳活动水平情况。并且，当耀斑发生时，能够及时给出预警信息并实时调整观测模式。该耀斑实时监测系统主要包括观测管理主系统、太阳活动水平评估模块以及耀斑捕捉定位模块三个部分。其基本架构如图 4.1 所示。

观测管理主系统包括太阳观测程序的主体部分以及其与“太阳活动水平评估模块”和“耀斑捕捉定位模块”两部分的信息交互管理部分。如图 4.2 是怀柔太阳观测系统的观测界面。该观测系统能够实时地采集CCD太阳图像并将其以一定的比例显示出来。除了常用的基本调节（如曝光时间、波带偏移和存盘间隔等参数的调节）外，该观测管理系统还有“常规模式”和“耀斑模式”两种观测模式可以选择。在“常规模式”下，该系统默认以60秒的时间间隔存储fits文件；而在“耀斑模式”下，可以用鼠标选取日面上的某一区域，然后以1秒的时间间隔存储fits文件，并且，在该模式下，系统将与“太阳活动水平评估模块”和“耀斑捕捉定位模块”有一定的信息交互，以实现太阳耀斑的智能化观测。“观测管理系统”将CCD 采集到的图像信息传送给“太阳活动水平评估模块”和“耀斑捕捉定位模块”，然后“太阳活动水平评估模块”和“耀

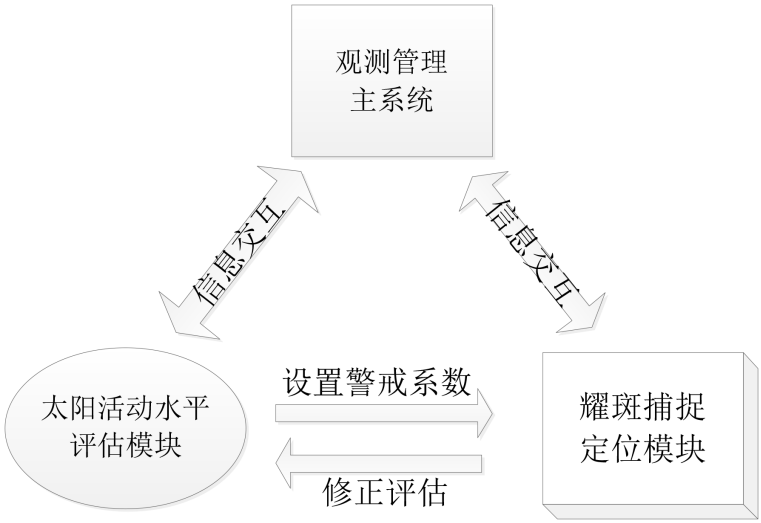


图 4.1: 太阳耀斑实时监测系统架构

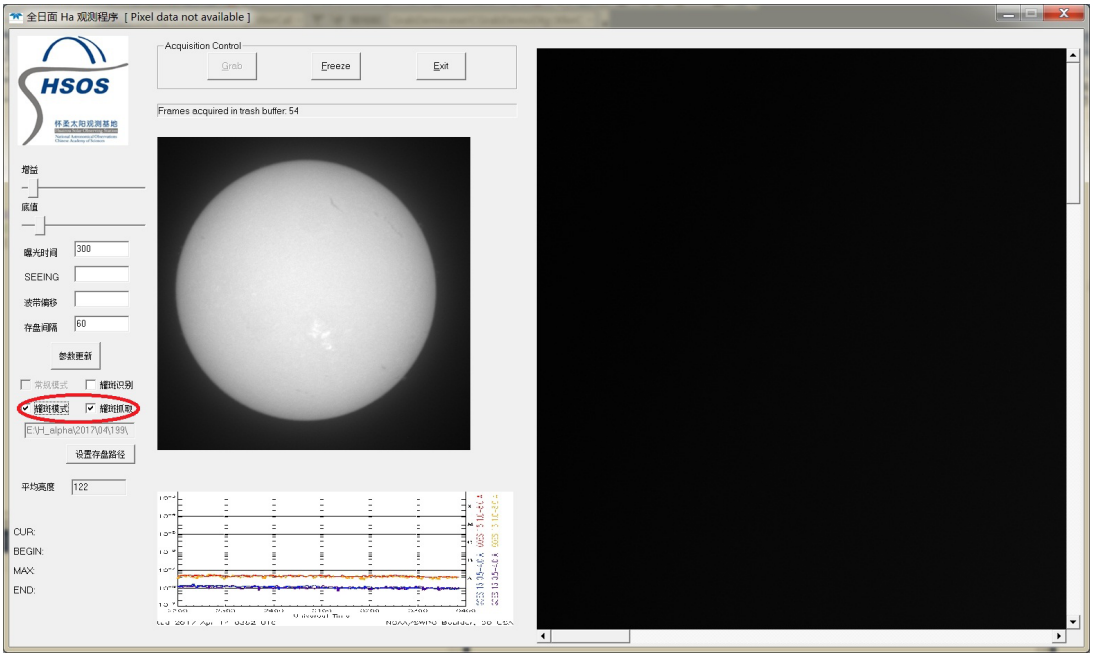


图 4.2: HSOS太阳观测系统监测界面

斑捕捉定位模块”在后台程序中分别评估太阳活动水平以及检测太阳耀斑的情况并将处理后的结果传送回“观测管理系统”。“观测管理系统”再以邮件、窗口提示的形式展示给观测人员。“太阳活动水平评估模块”和“耀斑捕捉定位模块”并非两个完全独立的模块，它们在“评估”、“捕捉”的同时也可以进行“交互”，以修正整个耀斑实时监测系统。

4.2　系统功能设计

“观测管理主系统”主要包括前台人机交互界面以及与另外两个部分的信息交互。在如图 4.2 所示的太阳观测界面中，观测人员可以调整各种设置以便于自己的观测，同时，“评估模块”和“耀斑捕捉定位模块”也是通过这一入口与观测人员进行交互的。当系统捕捉到一个耀斑时，将出现类似于：“怀柔太阳基地，智能化Hα耀斑实时监测系统，识别到一个可能的耀斑爆发，增亮区域面积331 （日面面积约4080744），爆发时间：2016-11-08 04:23:42，目前对应的GoesX 耀斑等级为B7.8，请关注。”所示的提醒信息。

“太阳活动水平评估模块”是在太阳耀斑预报模型基础上形成的评估模块，它的核心即为前文的PCA-SVM太阳耀斑预报模型。“太阳活动水平评估模块”的输入参量主要有：（1）“观测管理系统”CCD采集到的图像信息；（2）网络抓取数据；（3）本地计算机存储的全日面磁场数据。从“观测管理系统”CCD采集到的图像信息中，评估模块提取的特征包括： ① 灰度值的平均值； ② 灰度值标准差； ③ 与前一图平均值的差； ④ 重点监测点的亮度（key pixel，k=max(当前图灰度值-前图灰度值)）； ⑤ 计算出的k点径向位置； ⑥ k点亮度与7 × 7掩膜内最小亮度点的差值； ⑦ 以k 为中心的50 × 50 掩膜下的平均灰度值； ⑧ 50 × 50 掩膜内元素点的标准差； ⑨ 在对应掩膜下当前图灰度值与前一图灰度值的差值。从“网络抓取数据”中，评估模块提取的特征有射电流量数据（图 4.3 ）、前文所述的太阳黑子参量等数据。从“全日面磁场数据”中可提取的参量有： ① 太阳黑子个数、 ② 黑子面积、 ③ 最大黑子面积等。最后，PCA-SVM太阳耀斑预报模型将对这部分特征进行训练并输出评估后的结果。

“耀斑捕捉定位模块”主要是对“观测管理系统”CCD采集到的图像信息进行目标检测，包括日面的拟合、太阳活动特征的检测和最终耀斑的定位（图 4.4和 图 4.5 ）等。捕捉定位到的耀斑信息除了传输回“观测管理系统”交由其显示外，也便于“评估模块”的评估修正。同时，“耀斑捕捉定位模块”包括

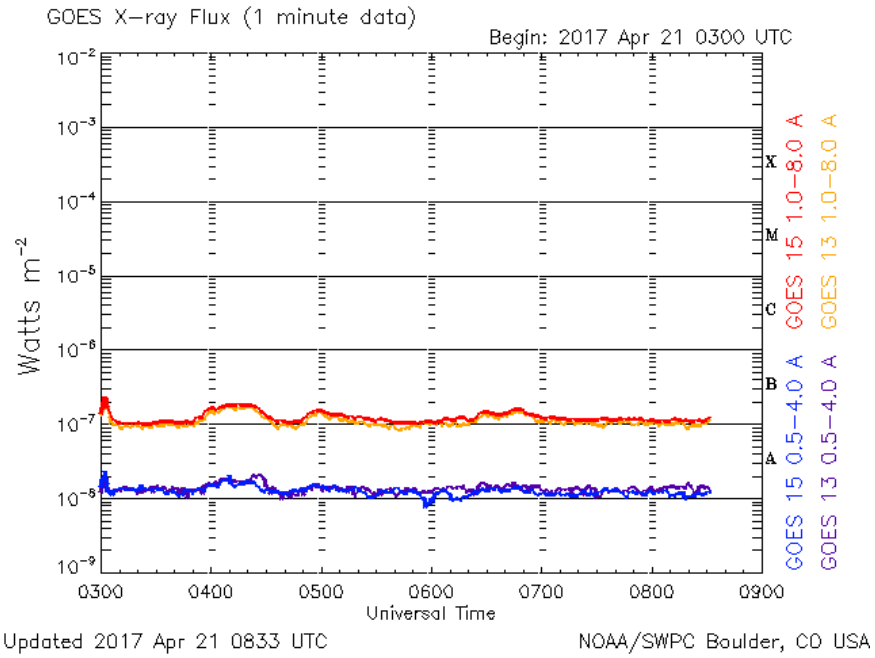


图 4.3: 耀斑实时监测系统抓取的射电流量数据

在GrabDemoDlg.cpp 下添加的耀斑监测功能（主要在函数CGrabDemoDlg::XferCallback 下添加），其基本过程如下：

1、图像预处理，对m_aDispBuf进行滤波去噪；

2、太阳日面拟合，确定太阳圆心和半径，以便于后期的太阳耀斑的定位；

3、判断是否存在耀斑（非偏带耀斑观测模式&&监测耀斑开始&&flearDetectByDivision ），其中，flearDetectByDivision的逻辑如下：

①、统计前一时刻保存的灰度值数组pre_buf[]以及现buf[]；

②、建立数组Division[]，统计现buf亮度较pre_buf的增长比；

③、设置一个阈值threshold1，将Division[]中大于这一阈值的区域标记为1，否则为0；

④、设置另一阈值threshold2，利用连通性算法合并部分③过程中为1 的区域，然后，只保留面积大于threshold2 的部分；

⑤、若Division 中不全为0，则返回true，否则返回false。

4、根据3 的返回值以及Division[]，标记出发生耀斑的区域。

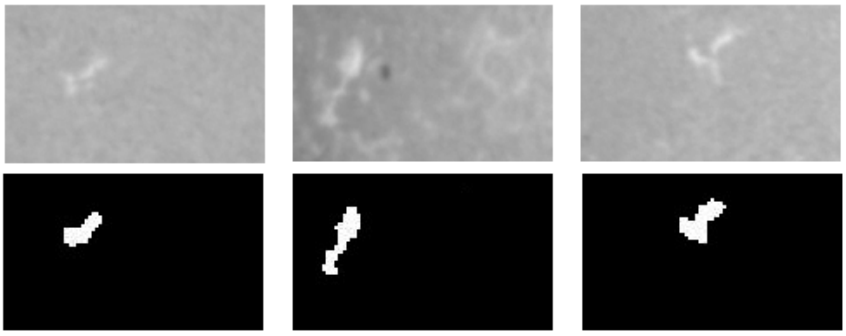


图 4.4: 太阳耀斑的检测

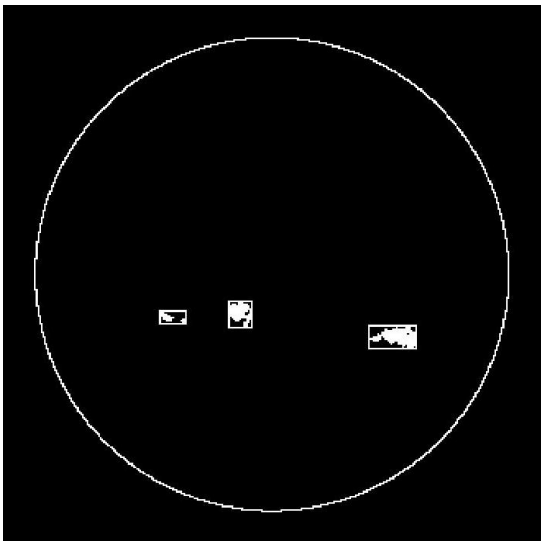


图 4.5: 太阳耀斑的捕捉定位

4.3　系统的实现

经过前期太阳活动水平评估模块中耀斑预报模型的阶段训练，本论文已完成了太阳耀斑实时监测系统的算法开发以及功能结构的设计。该智能化监测系统能够根据CCD采集数据以及网络抓取数据对太阳活动水平进行评估，当评估模块认为未来48小时内将发生M级及以上太阳耀斑时，它将以邮件形式提醒观测人员以提高其注意力。并且，在耀斑捕捉定位模块中，该系统能实时捕捉太阳耀斑的发生并将其定位，当选中“自动捕捉”选项时，主观测系统将实时修改观测参数（如曝光时间等参量），达到智能化观测的目的。目前该系统仍处于调试测试阶段，系统的稳定运行仍需要团队的进一步努力。

4.4 小结

在第二章“基于文本的预报”以及第三章“基于图像的检测”基础上，本章主要介绍了“基于全系统”的太阳耀斑实时智能化监测观测系统的具体研制工作。该实时智能化监测系统主要包含以下三个部分：“观测管理主系统”、“太阳活动水平评估模块”以及“耀斑捕捉定位模块”。其中，“基于文本的预报”可用于“太阳活动水平评估模块”，“基于图像的检测”可用于“耀斑捕捉定位模块”。本章节在宏观介绍这三个部分轮廓的同时，也一一介绍了它们的功能与具体实现方式。该监测系统一方面进行常规的太阳观测，另一方面，实时从太阳数据中提取出相关的预报因子，然后将这些预报因子嵌入太阳耀斑预报模型，以此评估太阳的活动水平，并实时捕捉、定位发生的太阳耀斑。

第五章　总结和展望

5.1　总结

太阳耀斑是一种剧烈的太阳活动现象，它的实时监测与预报不仅对太阳物理的科学研究具有重要的科研意义，而且对人类生产活动同样具有巨大的实用价值。为了进一步提高怀柔观测基地太阳耀斑实时监测与高分辨观测的水平和能力，本文将机器学习算法、图像处理技术和怀柔观测数据相结合，开展了太阳耀斑实时监测、预报的相关研究工作。本课题研究内容以及成果结论有以下四点。

第一，利用机器学习技术开展了分类、预测的研究工作，提出了一种新的基于机器学习方法的预报模型：PCA-SVM 预报模型。首先，对于输入的预报因子，通过协方差矩阵的计算、特征值的分解、特征向量的投影转换达到“降噪”和“去冗余”（即数据绛维）的目的，然后，经过分类算法的对比后采用泛化能力强、分类预测效果较好的支持向量机作为预报分类器。最后，核函数的选用有效避免了“维数灾难”且大大降低了分类预测模型的运算量。相较于多层感知器预报模型，PCA-SVM 模型具有运算复杂度低、泛化能力强等优点。

第二，将PCA-SVM预报模型用于太阳耀斑的预报工作，验证了PCA-SVM预报模型在太阳耀斑预报中的有效性。在这一部分工作中，重点介绍了太阳耀斑预报过程中（尤其是预报因子）的数据处理方法及其技术手段，太阳耀斑预报因子主要包括太阳黑子活动参量以及射电流量数据等。对于各种太阳耀斑的预报因子，有针对性地进行了“特征工程”流程，包括数据无量纲化、特征编码等数据的预处理以及后期的特征选择。最后，测试了2014年全年共2394 份的太阳耀斑预报的样本数据，结果验证了PCA-SVM 太阳耀斑预报模型的有效性。

第三，首次将引导滤波这一图像处理技术引入到天文图像的目标检测中，并提出了一种结合引导滤波和OTSU 阈值分割技术的太阳活动特征检测方法：GF-OTSU目标检测法。该方法结合了两者的优势并最终实现了太阳活动特征（包括太阳黑子、谱斑、耀斑等）的自动检测。本文以HSOS的2013年12月

共25天的太阳观测数据为测试数据对该方法进行检测测试，结果显示，该自动检测方法与人工检测结果具有较好的一致性。

第四，在前期“基于文本的预报”和“基于图像的检测”的基础上，完成了“基于系统”的怀柔观测基地太阳耀斑实时智能化监测系统的算法开发以及功能结构的设计工作。该实时智能化监测系统包括“太阳活动水平评估系统”、“太阳活动水平评估模块”和“耀斑捕捉定位模块”。“太阳活动水平评估模块”是基于PCA-SVM太阳耀斑预报模型而形成的太阳活动水平评估模块，“耀斑捕捉定位模块”用于对“观测管理主系统”中CCD 采集到的图像信息进行目标检测。这三个部分相互协同，共同构建出太阳耀斑的实时智能化监测系统。

5.2 展望

在太阳耀斑预报中，本文采用的预报因子主要为太阳黑子参量以及10.7cm射电流量，而研究表明，磁场剪切、磁螺度和电流螺度、纵向磁场最大水平梯度、中性线长度和孤立奇点等参量与太阳耀斑的发生也有着较高的相关性。下一步的工作可以尝试加入这些特征，以期构建出更为完善的太阳耀斑预报体系并获得更高的太阳耀斑预报准确率。另外，本文的PCA-SVM 太阳耀斑预报模型只预报了M级及以上太阳耀斑的有无，在后期的工作中可以考虑细分耀斑的级别，使该模型推广到多分类问题上。

对于文中的太阳活动特征检测方法，可以考虑将其推广至更多的太阳活动特征（如暗条、日珥等）检测中。在太阳耀斑实时智能化监测系统的研制中，目前该系统的协调能力还较为单一，未来可考虑引入更多的功能模块，进一步提升其太阳耀斑监测的“智能化”水平。

最后，希望本论文涉及的理论基础和相关技术积累未来可服务于正在进行的先进天基太阳天文台（ASO-S）项目。

参考文献

[1] 王华宁, 闫岩. 太阳活动的缓变与瞬变特征[J]. 气象科技进展, 2011, 01(4):59–61

[2] 林佳本. 高分辨太阳观测方法的研究[D]. 中国科学院国家天文台, 2009

[3] 王家龙, 孙静兰. 太阳活动及其对地球环境的影响[J]. 第四纪研究, 2002, 22(6):510–523

[4] 刘静远. 基于主成分分析对太阳周期活动及南北不对称性的研究[D]. 北京师范大学, 2010

[5] 苏江涛. 太阳矢量磁场测量[D]. 中国科学院国家天文台, 2004

[6] 胡新华. 日面活动特征的高速图像处理与识别系统研制[D]. 中国科学院国家天文台, 2007

[7] 林元章. 太阳物理导论[M]. 科学出版社, 2001

[8] 王璞. 太阳耀斑过程中的磁场变化研究[D]. 南京大学, 2011

[9] 方成, 丁明德, 陈鹏飞. 太阳活动区物理[M]. 南京大学出版社, 2008

[10] Schuurmans C J E. Influence of Solar Flare Particles on the General Circulation of the Atmosphere[J]. Nature, 1965, 205(4967):167–168

[11] 郭晓博, 王华宁, 戴幸华. 日冕物质抛射与太阳耀斑的时序关系分析[J]. 科学技术与工程, 2011, 11(13):2882–2887

[12] 王家龙, 张柏荣. 太阳活动预报简论[J]. 天文学进展, 1990, 8(2):89–98

[13] 焦维新. 空间天气学[M]. 气象出版社, 2003

[14] 崔延美. 太阳光球磁场特性与耀斑相关性研究[D]. 中国科学院研究生院(国家天文台), 2007

[15] LI R, ZHU J, HUANG X, et al. Solar flare forecasting model based on automatic feature extraction[J]. Chinese Science Bulletin, 2016, 61(36):3958–3963

[16] Mcintosh P S. The classification of sunspot groups[J]. Solar Physics, 1990, 125(2):251–267

[17] Bornmann P L, Shaw D. Flare rates and the McIntosh active-region classifications[J]. Solar Physics, 1994, 150(1):127–146

[18] Gallagher P T, Moon Y J, Wang H. Active-Region Monitoring and Flare Forecasting – I. Data Processing and First Results[J]. Solar Physics, 2002, 209(1):171–183

[19] Leka K D, Barnes G. Photospheric Magnetic Field Properties of Flaring versus Flare-quiet Active Regions. I. Data, General Approach, and Sample Results[J]. Astrophysical Journal, 2003, 595(2):pág̀s. 1277–1295

[20] Wheatland M S. A Bayesian Approach to Solar Flare Prediction[J]. Publications of the Astronomical Society of Australia, 2005, 609(2):153–156

[21] Bloomfield D S, Higgins P A, Mcateer R T J, et al. Toward Reliable Benchmarking of Solar Flare Forecasting Methods[J]. Astrophysical Journal Letters, 2012, 747(2):1112–1112

[22] Barnes G, Leka K D, Schumer E A, et al. Probabilistic forecasting of solar flares from vector magnetogram data[J]. Space Weather-the International Journal of Research Applications, 2016, 5(9):1–9

[23] Zhang G, Wang J, Li D. A new scheme used for the short-term prediction of X-ray flares.[J]. Publications of the Beijing Astronomical Observatory, 1994, 24

[24] Zhu C. Verification of the short-term prediction of solar X-ray bursts and solar proton events for the solar maximum (2000) of Solar Cycle 23[J]. Publications of the Beijing Astronomical Observatory, 2002, 38

[25] 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[J]. 计算机工程与设计, 2009, 30(15):3605–3607

[26] Wang H N, Cui Y M, Li R, et al. Solar flare forecasting model supported with artificial neural network techniques[J]. Advances in Space Research, 2008, 42(9):1464–1468

[27] Huang X, Yu D, Hu Q, et al. Short-Term Solar Flare Prediction Using Predictor Teams[J]. Solar Physics, 2010, 263(1):175–184

[28] Rong L, Wang H N, Cui Y M, et al. Solar flare forecasting using learning vector quantity and unsupervised clustering techniques[J]. Science China Physics, Mechanics Astronomy, 2011, 54(8):1546–1552

[29] 杨潇. 光球活动区磁非势性及其关联耀斑[D]. 中国科学院大学, 2013

[30] 申基, 胡柯良, 林佳本. 怀柔太阳观测基地三通道太阳望远镜局域网内远程观测终端系统设计[J]. 天文研究与技术, 2008, 5(3):281–287

[31] 艾国祥, 胡岳风. 太阳磁场望远镜的工作原理[J]. 天文学报, 1986, (2):91–98

[32] 赵翠. 基于MVC设计模式的怀柔数据查询系统[D]. 中国科学院研究生院, 2011

[33] Mitchell T M, Carbonell J G, Michalski R S. Machine Learning[M]. China Machine Press, 2003: 417–433

[34] Bache K, Lichman M. UCI Machine Learning Repository[J]. 2013.

[35] Dietterich T G. Machine-Learning Research; Four Current Directions[M]. 1997

[36] 何清, 李宁, 罗文娟. 大数据下的机器学习算法综述[J]. 模式识别与人工智能, 2014, 27(4):327–336

[37] Liu S, Dong L, Zhang J, et al. 深度学习在自然语言处理中的应用[J]. 中国计算机学会通讯, 2015.

[38] Michie D, Spiegelhalter D J, Taylor C C, et al. Machine Learning, Neural and Statistical Classification[J]. Technometrics, 1995, 37(4):459–459

[39] Hastie T, Tibshirani R, Friedman J. Springer Series in Statistics[M]. 2009: 1–17

[40] SáJ P M D. Statistical Classification[M]. Betascript Publishing, 2001: 231–246

[41] Bland J M, Altman D J. REGRESSION ANALYSIS[J]. Lancet, 1986, 1(8486):908–909

[42] Jain A K, Murty M N, Flynn P J. Data clustering: a review[J]. Acm Computing Surveys, 1999, 31(3):264–323

[43] Imielienskin T, Swami A, Agrawal R. Mining association rules between set of items in largedatabases[J]. Acm Sigmod Record, 1993, 22(2)

[44] IanHWitten, EibeFrank. 数据挖掘实用机器学习技术[M]. 机械工业出版社, 2006

[45] 李蓉, 崔延美, 贺晗. 基于支持向量机和k近邻的太阳质子事件预报模型[J]. 科学技术与工程, 2007, 7(15):3649–3654

[46] Ali S, Smith K A. On learning algorithm selection for classification[M]. Elsevier Science Publishers B. V., 2006: 119–138

[47] Quinlan J R. Induction on decision tree[J]. Machine Learning, 1986, 1(1):81–106

[48] Rish I. An empirical study of the naive Bayes classifier[J]. Journal of Universal Computer Science, 2001, 1(2):127

[49] Bishop C M. Neural Networks for Pattern Recognition[J]. Agricultural Engineering International the Cigr Journal of Scientific Research & Development Manuscript Pm, 1995, 92(440):1642

[50] Haddadi F, Khanchi S, Shetabi M, et al. Intrusion Detection and Attack Classification Using Feed-Forward Neural Network[M]. 2010: 262–266

[51] Hanson S J, Pratt L. Comparing biases for minimal network construction with back-propagation[M]. Morgan Kaufmann Publishers Inc., 1989: 177–185

[52] Zhang Z Y. Stochastic Neural Network[M]. Springer New York, 2013: 1998–1998

[53] Cohen M A, Grossberg S. Absolute Stability of Global Pattern Formation and Parallel Memory Storage by Competitive Neural Networks[J]. IEEE Transactions on Systems Man & Cybernetics, 1970, SMC-13(5):815–826

[54] Guo G, Wang H, Bell D, et al. KNN Model-Based Approach in Classification[M]. Springer Berlin Heidelberg, 2003: 986–996

[55] Danielsson P E. Euclidean distance mapping[J]. Computer Graphics & Image Processing, 1980, 14(3):227–248

[56] Craw S. Manhattan Distance[M]. Springer US, 2011: 639–639

[57] Surhone L M, Tennoe M T, Henssonow S F. Chebyshev Distance[M]. 2010

[58] Liu C. Discriminant analysis and similarity measure[J]. Pattern Recognition, 2014, 47(1):359–367

[59] 温靖. 太阳活动指数的中期预报方法研究[J]. 2009.

[60] 王华宁. 太阳活动综合预报系统[J]. 2007.

[61] 李蓉. 人工智能在太阳活动预报中的应用[D]. 中国科学院研究生院, 2007

[62] 李希灿, 邱发堂. 预报因子选择的模糊优选方法[J]. 预测, 1999, (4):78–80

[63] 袁飞, 林佳本, 邓元勇. 结合主成分分析和支持向量机的太阳耀斑预报模型[J]. 科学通报, 2016, (20):2316–2321

[64] Dubrova E. Information Redundancy[M]. Springer New York, 2013: 87–136

[65] Turner C R, Wolf A L, Fuggetta A, et al. Feature Engineering[J]. Proceedings of International Workshop on Software Specification & Design, 1998. 162–164

[66] Leigh D A, Alessro T, Francesco Z. Reducing Molecular Shuttling to a Single Dimension.[J]. Angewandte Chemie International Edition, 2000, 39(2):350

[67] Deutsch H P. Principle Component Analysis[J]. 2002.

[68] Deutsch H P. Principle Component Analysis[M]. Palgrave Macmillan UK, 2004

[69] Brockett P L, Xia X, Derrig R A. Using Kohonen's Self-Organizing Feature Map to Uncover Automobile Bodily Injury Claims Fraud[J]. Journal of Risk & Insurance, 1998, 65(2):245–274

[70] Cambria E, Mazzocco T, Hussain A. Application of multi-dimensional scaling and artificial neural networks for biologically inspired opinion mining[J]. Biologically Inspired Cognitive Architectures, 2013, 4(4):41–53

[71] 宗序平, 姚玉兰. 利用Q-Q图与P-P图快速检验数据的统计分布[J]. 统计与决策, 2010, (20):151–152

[72] Allen D. Automatic one-hot re-encoding for FPGAs[J]. Proceedings of the International Workshop on Field Programmable Logic & Applications, 1992, 705:71–77

[73] Pissanetzky S. Sparse matrix technology /[M]. Academic Pr, 1984: 1 – 3

[74] Anis M Z. Determination of the Best Mean Fill[J]. Quality Engineering, 2003, 15(3):407–409

[75] Jian-Xiong L, Xin L I, Cheng D W. The application of linear regression method in nuclear data processing[J]. Journal of Northeast Normal University, 2014.

[76] Cui B, Hu J, Shen H, et al. Adaptive Quantization of the High-Dimensional Data for Efficient KNN Processing[M]. Springer Berlin Heidelberg, 2004: 302–313

[77] Yang Y. Robust bayesian estimation[J]. Journal of Geodesy, 1991, 65(3):145–150

[78] Ukil A. Support Vector Machine[J]. Computer Science, 2002, 1(4):1–28

[79] 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[D]. 2013

[80] 陈哲, 王慧斌. 图像目标检测技术及应用[M]. 人民邮电出版社, 2016

[81] Jain A K. Fundamentals of digital image processing[J]. 2010, 16(10):1420–1424

[82] Gauch J M. Image segmentation and analysis via multiscale gradient watershed hierarchies[M]. IEEE Press, 1999: 69 – 79

[83] Comer M L, Delp E J. Morphological operations[M]. Springer US, 1998: 210–227

[84] Agam G, Dinstein I. Regulated morphological operations ☆[J]. Pattern Recognition, 1999, 32(6):947–971

[85] Otsu N. Threshold Selection Method from Gray-Level Histograms, IEEE Transactions on Systems Man and Cybernetics[J]. Systems Man & Cybernetics IEEE Transactions on, 1979, 9(1):62–66

[86] Otsu N. An automatic threshold selection method based on discriminate and least squares criteria[J]. Denshi Tsushin Gakkai Ronbunshi, 1979, 63:349–356

[87] He K, Sun J, Tang X. Guided Image Filtering[M]. Springer Berlin Heidelberg, 2010

[88] Pablo B, Martin E, Marcus M. Guided Image Filtering for Interactive HighGlobal Illumination[J]. Computer Graphics Forum, 2011, 30(4):1361–1368

[89] 林佳本, 沈洋斌, 朱晓明. 怀柔太阳观测基地全日面磁场自动化观测系统[J]. 天文研究与技术:国家天文台台刊, 2013, 10(4):392–396

发表文章目录

[1] **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[J]. 科学通报, 2016(20):2316-2321.

[2] **Fei Yuan**, Jiaben Lin, Jingjing Guo, et al. Automatic detection of solar features in HSOS full-disk solar images using guided filter[J]. New Astronomy.

简　　历

基本情况

袁飞，男，汉族，中国科学院国家天文台在读硕士研究生。

教育状况

2014 年 9 月至今，中国科学院大学，中国科学院国家天文台，硕士，专业：天文技术与方法；

2010 年 9 月至 2014 年 7 月，北京航空航天大学，自动化科学与电气工程学院，本科，专业：自动化。

获奖情况

2015 至 2016 学年，中国科学院国家天文台AMD奖学金一等奖；

2011 至 2012 学年，教育部 国家励志奖学金。

研究兴趣

机器学习、图像处理、太阳活动监测、太阳耀斑预报

联系方式

通讯地址：北京市朝阳区大屯路甲 20 号 中国科学院国家天文台

邮编：100012

电子邮箱：yfei34@163.com

致　　谢

时光荏苒，转眼间我的研究生生活也将结束了。回顾三年的研究生生活，在中国科学院大学、国家天文台的这三年，于我而言是一份宝贵的人生财富。三年里，有欣慰有遗憾，有酸甜有苦辣，但随着研究生生涯的结束，这一切也将尘埃落定。在此毕业之际，我谨向所有关心、支持和帮助过我的老师、同学、朋友和家人们表示诚挚的感谢以及美好的祝愿。

首先，衷心感谢我的导师林佳本老师，感谢林老师对我的谆谆教诲，感谢他在我学习、工作、生活里无微不至的悉心关怀和照顾。林老师严谨的治学态度和温和的为人处世原则，默默浸润着我的学习工作和生活。在论文的撰写过程中，从论文理论基础、到数据的搜集、论文的架构乃至无数细节之处的修改，林老师无不悉心指导。论文之内，我有成长，论文之外，我更是受益匪浅。在此，向导师表示由衷的感谢，谢谢您林老师，感谢您将我领入天文技术研究的道路。

感谢怀柔基地的所有的成员，感谢这个温馨的家庭，感谢邓元勇老师、张洪起老师、王东光老师在百忙之中尽职尽责地指导我们的科研工作。感谢组里的其他老师，几位老师知识渊博、科研经验丰富、科研认真负责，对我的帮助很大。感谢研究生处的杜红荣老师、艾华老师、马怀宇老师对我的帮助。感谢白先勇师兄、杨潇师姐、赵翠师姐在我论文的撰写中提出的很多宝贵意见。三年研究生生活，实验室里共同的生活点滴，感谢郭晶晶师姐、王刚同学、曾祥云同学以及在实验室内陪我共同走过这三年的其他同学们，是你们的陪伴让我的研究生生活变得绚丽多彩。

最后，再次感谢所有在本文的撰写和学习工作中给予我关怀和帮助的老师和同学们，祝他们身体健康、事业有成、生活幸福美满!