

随机森林算法实现小行星分类*

黄超^{1,2,3} 马月华^{1,2†} 赵海斌^{1,2,4} 卢晓平⁵

(1 中国科学院紫金山天文台 南京 210008)

(2 中国科学院行星科学重点实验室 南京 210008)

(3 中国科学院大学 北京 100049)

(4 澳门科技大学月球与行星科学实验室—中国科学院月球与深空探测重点实验室伙伴实验室
澳门 000853)

(5 澳门科技大学资讯科技学院 澳门 000853)

摘要 随着小行星光谱和测光数据不断增加, 以此为基础已有多种分类方法对小行星进行分类. 使用随机森林算法对Sloan Digital Sky Survey (SDSS) Moving Object Catalogue (MOC) 的观测结果进行小行星分类. 根据小行星g、r、i、z 4个波段的星等, 结合Tholen、Bus、Lazzaro、DeMeo等人已有的分类工作和主成分分析, 对多达48642颗的小行星进行了分类, 实现了大数据集的小行星类别判定, 把这些小行星分为8个类别(C、X、S、B、D、K、L 和V).

关键词 小行星: 普通, 谱线: 轮廓, 方法: 随机森林算法
中图分类号: P185; 文献标识码: A

1 引言

小行星是太阳系中重要的小天体, 对其观测研究已经成为当代天文学中的一个活跃领域. 获得小行星确切的组成成分, 能更加深入地了解在太阳系形成时期的环境条件及演化过程. 太阳系中地球、火星等类地行星, 由于其本身的演化过程和大气环境等因素的影响, 已经很难在此类行星上发现太阳系早期形成所留下的证据, 而小行星作为类地行星形成时期的组成成分, 特别是处于主带和特洛伊带的小行星, 依然保持着数十亿年前的状态, 留下了太阳系早期形成时期的重要证据^[1]. 小行星测光数据分析是研究小行星组成成分的一种重要手段, 也为小行星分类方法的研究提供了一个重要的依据.

1929年, Bobrovnikoff^[2]发表了对小行星光谱测量的研究报告, 但是受当时技术发展的限制, 光谱测量效率低下且不够精确. 20世纪50年代中期, 随着UBV光度测量的使用, 研究人员拥有了大量小行星的分光观测资料. 利用这些资料, Wood等^[3]、Chapman等^[4]

2016-03-28收到原稿, 2016-04-22收到修改稿

*国家自然科学基金项目 (11573075, 11403107, 11273067)、江苏省自然科学基金项目 (BK20141045)、澳门科学技术发展基金项目 (095/2013/A3)、紫金山天文台小行星基金会及澳门科技大学月球与行星科学实验室—中国科学院月球与深空探测重点实验室伙伴实验室资助

†yhma@pmo.ac.cn

获得了小行星光谱的反射特性, 并以此把小行星分为了两个大类, 一类即为“S”类小行星, 另一类为“C”类小行星. Zellner^[5]发现小行星反射率有着双峰分布的特点, 由此把小行星分为暗碳质和亮石质两个大类. 随后几年, 越来越多的科研项目投入到了小行星光学特性的研究中, 为20世纪70年代中期提出一种更为精确且严格的小行星分类法提供了可能性. Chapman等^[6]在分析了一些小行星光度测量的反照率后, 进一步把小行星细分为“C”、“S”、“U”3类: “C”代表暗的碳质类; “S”代表具有石质类特性; “U”代表还不能确定类别. 不同类别小行星在一定意义上揭示了其物质组成成分的不同.

我们结合已有的分类结果, 用随机森林算法对Sloan Digital Sky Survey (SDSS) Moving Object Catalogue (MOC)¹观测的大部分小行星进行类别判定. SDSS使用美国新墨西哥州阿帕奇天文台的2.5 m口径望远镜作为观测设备, 望远镜的测光系统使用了多个波段滤光片, 分别为u、g、r、i、z波段, 中心波长分别为: 3551 Å、4686 Å、6166 Å、7480 Å 和8932 Å^[7]. 该巡天项目观测成果丰硕, 得到了大量高精度的观测数据. 叶嘉晖等^[8-9]利用主成分分析法对SDSS的数据进行了小行星分类, 有效区分了S大类和C大类小行星反射率特点, 但受主成分分析方法的限制, 得到的数据分辨率不高, 只把小行星分作了两个大类. 随机森林算法能克服该方法的局限性, 能以更高精度识别出不同小行星反射率的特点, 从而分出多个类别.

2 随机森林算法

传统的分类模型存在容易出现过拟合、精度不高等缺点^[10], 很多研究者开始集合多个分类模型来提高分类的精度. 根据训练集创建一组基分类模型, 然后对每个基分类模型得到的预测值进行投票来最终决定预测值.

随机森林算法正是这样一种包含一组基分类模型的算法, 利用bootstrap重抽样方法在原始样本中抽取若干样本, 对每个样本都进行决策树建模. 通过训练集的训练, 每棵决策树产生一个类别的判定准则. 对于每一个需要判定类别的小行星, 每棵决策树会产生一个分类结果, 最终, 采用多数表决的方法作为预测结果输出.

随机森林优越的分类性能使其在学术研究中得到了广泛的应用, 我们使用R语言软件包randomForest来实现. 通过训练集的训练, 随机森林能识别出训练集中不同类别小行星反射光谱特点, 对比未知类别小行星的反射光谱, 实现分类.

2.1 数据处理

在SDSS MOC观测数据中, 去除了有较大观测误差的个体, 且只选取已得到编号的小行星, 这样得到了48642颗小行星. 对于一颗小行星, 我们用g、r、i、z、v 5个波段的星等求得色指数(由于u波段的星等数据相对于其他波段误差显著过大, 参照以往对SDSS数据进行小行星分类的方法, 采用了SDSS MOC数据中提供的v波段星等, v波段

¹<http://sbn.psi.edu/pds/resource/sdssmoc.html>

的星等由g和r波段的星等计算得出). 由以下色指数计算公式^[11], 求得:

$$C_{g-v} = g - v - 0.2720, \quad (1)$$

$$C_{v-r} = v - r - 0.1780, \quad (2)$$

$$C_{v-i} = v - i - 0.2780, \quad (3)$$

$$C_{v-z} = v - z - 0.3180, \quad (4)$$

上式中g、r、i、z、v分别指该波段星等, C_{g-v} 、 C_{v-r} 、 C_{v-i} 、 C_{v-z} 分别为计算得到的色指数.

在计算相对反射率时, 利用Misra等人的结果^[11], 小行星相对反射率的计算公式为:

$$F_g = 10^{-0.4C_{g-v}}, \quad (5)$$

$$F_r = 10^{0.4C_{v-r}}, \quad (6)$$

$$F_i = 10^{0.4C_{v-i}}, \quad (7)$$

$$F_z = 10^{0.4C_{v-z}}, \quad (8)$$

F_g 、 F_r 、 F_i 、 F_z 分别为g、r、i、z波段的相对反射率. 我们利用以上式子得到的相对反射率对小行星进行分类.

2.2 训练集的选取

在选取训练集的过程中, 参考了Tholen^[12]、Bus等^[13]、Lazzaro等^[14]、DeMeo等^[15]的分类结果, 在SDSS的数据中找出了210颗已有分类结果的小行星. 但是有些类别的小行星只根据SDSS的星等数据不能得到区分: 比如在Bus分类的26个类别中, C和Ch类在0.7 μm 处的吸收峰特征有所不同, 然而这些特点在SDSS数据中得不到体现, 我们不能依据SDSS的数据区分出C和Ch类. 因此, 我们把这210颗小行星分作了8个类别, 这8个类别与Bus分类结果的对应关系见表1.

要对48642颗小行星分类, 如果只以这210颗小行星作为训练集, 数量太少, 不能得到一个预期的分类结果, 因此需要扩充训练集的小行星数量. 小行星分类的依据是相对反射率, 就是把相对反射率相似的小行星分为一类. 我们使用主成分分析法, 选取相对反射率与已知分类小行星相似的小行星, 判定其类别.

主成分分析为数学变换方法的一种, 把相关变量转换成互不相关的变量, 转换后的变量以方差大小排列, 方差最大的作为第1个变量, 命名为第1主成分(PC1), 方差次大的作为第2个变量, 命名为第2主成分(PC2), 以此类推. 方差越大, 变量包含的信息越多, 因此, 由主成分分析得到的第1主成分包含原变量的主要信息. 通过主成分分析, 我们就能以少数几个主成分变量替代原先多个变量, 实现降维.

在小行星色指数研究中, 计算主成分的公式为:

$$PC = E(C - B), \quad (9)$$

PC为主成分, C为色指数, B是色指数均值, E是色指数协方差矩阵的特征向量. 由此得到的第1主成分反映了反射光谱的颜色特征.

表 1 本文的分类与Bus分类的对应关系

Table 1 Our taxonomic system compared to the Bus taxonomic system

Our Taxonomy	Bus Taxonomy
B	B, Cb
C	C, Cb, Cg, Ch, Cgh
D	D, T
L	L, Ld, Sl
K	K, T, Sk, Xk
S	S, A, Q, R, Sa, Sk, Sl, Sq, Sr
X	X, Xc, Xk, Xe
V	V, O, Q, R

对48642颗小行星进行主成分分析后, 我们得到的第1主成分和第2主成分包含了95.98%的相对反照率信息(见表2), 第1主成分和第2主成分基本就代表了相对反照率的全部信息, 并以此作出图1, 从图1中我们可以明显地看出小行星双峰分布的特点.

表 2 本文计算得到的主成分包含信息百分比

Table 2 Percentage of variance accounted by the principal component

	PC1	PC2	PC3	PC4
Variance	0.1313	0.0858	0.0318	0.0043
Percentage	67.27%	28.71%	3.95%	0.07%
Cumulative Percentage	67.27%	95.98%	99.93%	100.00%

结合得到的第1、第2主成分, 选取出PC1和PC2与上文已知分类小行星相近的个体, 并结合它们相对反射率的特点, 我们又选出910颗小行星, 进行了类别的判定(训练集中小行星在PC空间的分布见图2). 以这样的方式, 我们把训练集扩充到了1120颗.

3 结果与分析

在以训练集训练分类模型时, 我们尝试选取了500到1000棵决策树的随机森林进行训练, 发现决策树超过600棵后分类模型就不再显著地改进, 我们把决策树设定为600棵. 训练集训练好模型后, 我们得到了一个随机森林的分类模型, 就可以对余下的绝大部分小行星进行分类. 得到的分类结果见表3和图3.

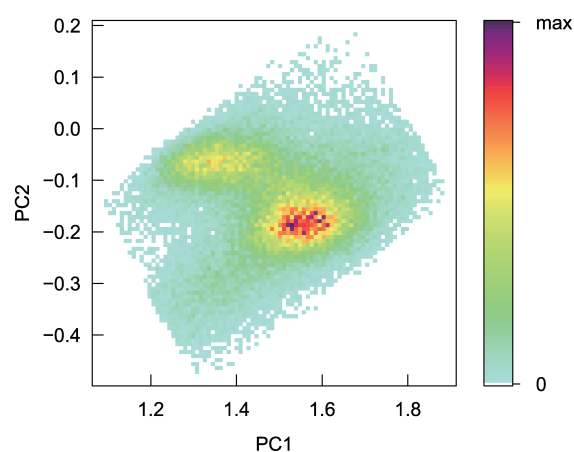


图 1 通过PC1和PC2空间分布图能看到明显的双峰结构

Fig. 1 The bimodal structure of asteroids shown in the PC space

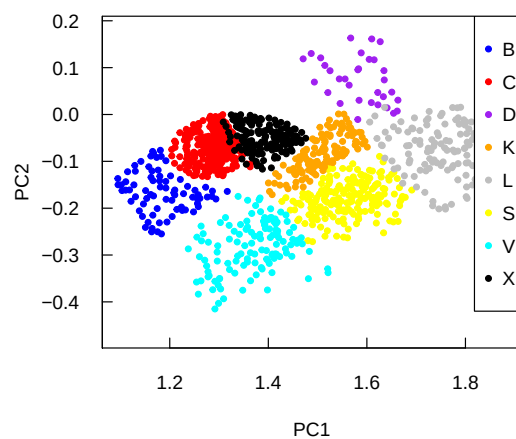


图 2 训练集中的小行星在PC空间的分布

Fig. 2 The distribution of asteroids of the training sets in the PC space

表 3 各类别小行星所含数量

Table 3 Percentages of asteroids classified into each asteroid type

	B	C	D	K	L	S	V	X
Number	630	6250	990	4418	3747	20747	5531	6329
Percentage	1.30%	12.85%	2.04%	9.08%	7.70%	42.65%	11.37%	13.01%

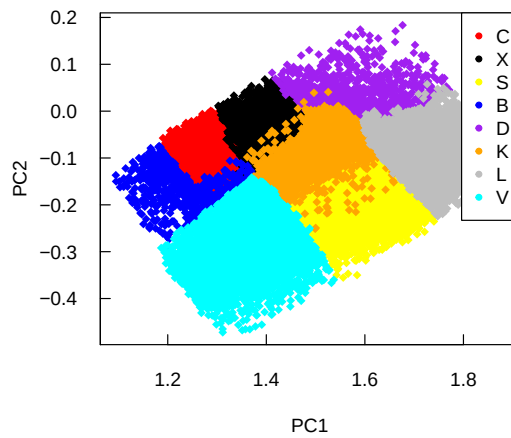


图3 分类结果在PC空间中的分布, 从中可以看到每个类别的小行星都聚集成类

Fig. 3 The distribution of different classes of asteroids in the PC space, in which each class of asteroids are clustered

为了与已有的小行星分类结果进行对比, 我们根据每个类别小行星在各个波段反射率的均值, 做出了小行星的反射光谱图, 见图4. 可以看到, 每个类别的小行星反射光谱与已有分类结果的小行星反射光谱特点基本保持一致: S类小行星反射光谱有明显的波峰、C和X类小行星反射光谱比较平坦等, 表明我们的分类取得了较好的结果.

4 总结

基于g, r, i和z波段获得的星等数据, 用随机森林算法实现了大数据量的小行星分类. 通过得到的分类结果, 可以看出随机森林算法准确地分辨了每个类别小行星相对反射率的特点.

但是, 有些类别的分类结果还需要后续的改进. 比如K类小行星, 由于训练集中该类别小行星数量较少, 且与S和X类小行星相对反射率特点区分度不高, 随机森林模型对于该类别的小行星分类结果有一定的弥散.

对于大量小行星观测数据, 通过主成分分析法, 在PC空间中划分区域来实现小行星分类的方法效率不高, 随机森林算法能高效率地实现大数据量小行星的分类. 随着小行星观测数据量的不断增加, 随机森林算法的快速分类应用价值更能得到体现.

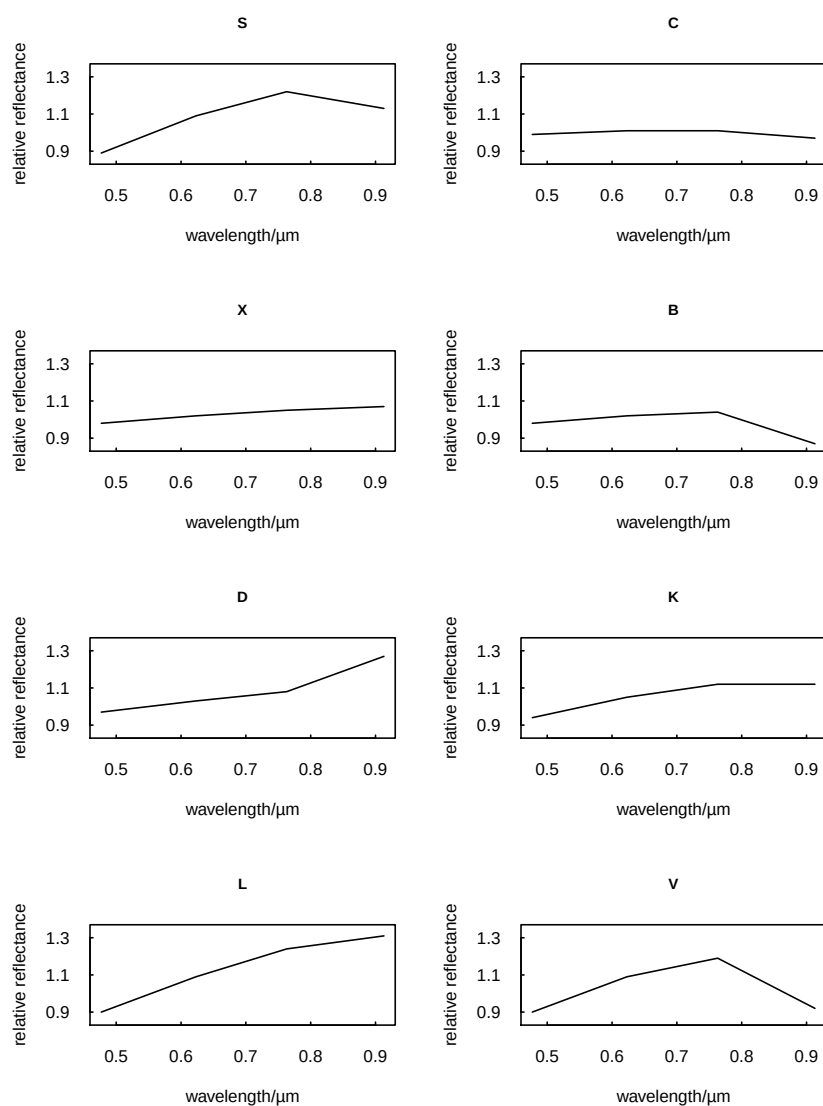


图 4 8个类别小行星相对反射光谱图

Fig. 4 The relative reflectance spectra of the eight classes of asteroids

参 考 文 献

- [1] Gaffey M J, Burbine T H, Piatek J L, et al. Icarus, 1993, 106: 573
- [2] Bobrovnikoff N T. Lick Observatory Bulletin, 1929, 14: 18
- [3] Wood J H, Kuiper G P. ApJ, 1963, 137: 1279
- [4] Chapman C R, Johnson T V, McCord T B. NASA Special Publication, 1971, 267: 51
- [5] Zellner B. Bulletin of the American Astronomical Society, 1973, 5: 388
- [6] Chapman C R, Morrison D, Zellner B. Icarus, 1975, 25: 104
- [7] Ivezić Ž, Tabachnik S, Rafikov R, et al. AJ, 2001, 122: 2479
- [8] 叶嘉晖, 赵海斌, 李彬. 天文学报, 2015, 56: 243

- [9] Ye J H, Zhao H B, Li B. ChA&A, 2016, 40: 54
- [10] 方匡南, 吴见彬, 朱建平, 等. 统计与信息论坛, 2011, 26: 32
- [11] Misra A, Bus S J. BAAS, 2008, 40: 508
- [12] Tholen D J. Asteroid Taxonomic Classifications. Tucson: University of Arizona Press, 1989
- [13] Bus S J, Binzel R P. Icarus, 2002, 158: 146
- [14] Lazzaro D, Angeli C A, Carvano J M, et al. Icarus, 2004, 172: 179
- [15] DeMeo F E, Binzel R P, Slivan S M, et al. Icarus, 2009, 202: 160

Spectral Classification of Asteroids by Random Forest

HUANG Chao^{1,2,3} MA Yue-hua^{1,2} ZHAO Hai-bin^{1,4} LU Xiao-ping⁵

(1 Purple Mountain Observatory, Chinese Academy of Sciences, Nanjing 210008)

(2 Key Laboratory for Planetary Science, Chinese Academy of Sciences, Nanjing 210008)

(3 University of Chinese Academy of Sciences, Beijing 100049)

(4 Lunar and Planetary Science Laboratory, Macau University of Science and Technology—Partner Laboratory of Key Laboratory of Lunar and Deep Space Exploration, Chinese Academy of Sciences, Macau 000853)

(5 Faculty of Information Technology, Macau University of Science and Technology, Macau 000853)

ABSTRACT With the increasing asteroid spectral and photometric data, a variety of classification methods for asteroids have been proposed. This paper classifies asteroids based on the observations of Sloan Digital Sky Survey (SDSS) Moving Object Catalogue (MOC) by using the random forest algorithm. With the training data derived from the taxonomies of Tholen, Bus, Lazzaro, DeMeo, and Principal Component Analysis, we classify 48642 asteroids according to g, r, i, and z SDSS magnitudes. In this way, asteroids are divided into 8 spectral classes (C, X, S, B, D, K, L, and V).

Key words asteroids: general, line: profiles, methods: random forest