低信噪比巡天数据中特殊恒星光谱的搜寻方法

吴明磊[1]，潘景昌[1*]，衣振萍[1]，韦　鹏[2]

1. 山东大学(威海)机电与信息工程学院，山东 威海　264209
2. 中国科学院光学天文重点实验室，国家天文台，北京　100012

摘　要　特殊恒星是金属丰度异常的恒星，其中包含的信息对于研究宇宙起源、太阳系的演变以及生命的演化都有着重要的意义。因此，特殊恒星的搜寻是国内外巡天项目中的重要目标。恒星光谱中包含着恒星的化学成分、物理性质以及运动状态等丰富的信息，它是开展恒星研究的重要依据。恒星的识别、分类以及特殊恒星的发现主要依据的是恒星光谱数据。随着LAMOST和SDSS等国内外大规模数字巡天项目的深入展开，恒星光谱的数据量达到了前所未有的高度，如此大的数据量为特殊恒星的发现提供了强有力的支撑。因此如何利用这些数据快速准确地发现特殊、稀少甚至于未知类型的恒星光谱是天文学研究的重要问题。数据挖掘是结合模式识别、机器学习、统计分析及相关专家背景知识，从数据中提取出隐含的过去未知的有价值的潜在信息的技术，其在处理大数据方面有着天然的优势，越来越多的数据挖掘方法被应用到巡天数据处理及分析之中。目前针对特殊恒星搜寻的数据挖掘算法主要包含随机森林、聚类分析以及异常值检测等，但随着巡天深度的拓展，观测的目标越来越暗，进而观测光谱的信噪比也随之变低。低信噪比光谱中存在着大量的无用信息，直接利用相关算法对其进行分析处理得到的结果往往存在很大的偏差。因此，如何从大量低信噪比恒星光谱巡天数据中有效地搜寻出特殊的恒星光谱，是当前面临的一个重要问题。由于低信噪比恒星光谱本身的特点，对于从中搜寻特殊恒星光谱的工作开展较少。为了解决此问题，在仔细研究光谱数据处理方法的基础上，针对低信噪比巡天数据中特殊恒星光谱的搜寻，提出了一种以主成分分析(PCA)和基于密度峰值聚类为基础的方法。该方法首先选取O，B，A，F，G，K和M各种类型的高信噪比恒星光谱，进行波长统一和流量插值后，利用主成分分析得到特征光谱；然后利用方差贡献率最大的前几个特征光谱对低信噪比的恒星光谱进行重构得到高信噪比的光谱；最后利用重构之后的高信噪比光谱进行聚类，聚类分析中得到的离群数据即为所要搜寻的特殊恒星光谱。在聚类时，考虑到恒星光谱数据本身的特点，采用了一种基于密度峰值的聚类方法来进行聚类及离群点的挖掘。实验表明，该方法能够在低信噪比的恒星光谱巡天数据中准确地搜寻出数量相对较少的特殊恒星。同时，也可应用于诸如LAMOST、SDSS等各种银河系巡天的光谱数据分析与挖掘中。

关键词　银河系巡天；离群数据挖掘；低信噪比光谱

中图分类号：P145.4　　**文献标识码**：A　　**DOI**：10.3964/j.issn.1000-0593(2019)02-0618-04

引　言

多目标光谱技术可以同时观测成百上千个目标，观测的效率得到极大的提高。因此，大规模巡天项目将获得海量光谱数据[1-2]。如此巨大的数据不仅为银河系及河外科学研究提供了大量研究样本，还为发现一些异常、稀少甚至于未知类型的观测目标提供了可能性[3]。随着巡天深度的拓展，需要观测的目标越来越暗，同样的观测条件下获得光谱的信噪比也随之变低。如何从这些低质量低信噪比光谱中搜寻其中的特殊恒星光谱，是当前天文研究中的一个重点和难点。

随着大规模巡天项目的不断进行，越来越多的数据挖掘方法应用到巡天数据处理及分析之中。目前国内外已经开展了相关研究工作，比如利用神经网络、SVM及标签传递等算法进行恒星的分类、搜寻及特殊天体的查找[4-8]。

可以看出，由于低信噪比恒星光谱本身的特点，对于从

收稿日期：2018-01-22，**修订日期**：2018-05-18

基金项目：国家自然科学基金项目(U1431102，11603014)资助

作者简介：吴明磊，1986年生，山东大学(威海)机电与信息工程学院博士研究生　　e-mail：wuming8511@126.com

*通讯联系人　　e-mail：pjc@sdu.edu.cn

中搜寻特殊恒星光谱的工作开展较少。本文在已有搜寻特殊类型光谱方法的基础上，提出一种从低信噪比巡天光谱数据中搜寻特殊恒星光谱的方法。该方法利用高信噪比恒星光谱进行主成分分析得的特征光谱，对低信噪比的恒星光谱进行重构，然后对进行重构之后的光谱进行聚类，聚类分析中得到的离群数据即为所要搜寻的特殊恒星光谱。实验表明，本文提出的方法在低信噪比的光谱数据中能有效的发现其中数量较少的特殊恒星光谱。

1　方法介绍

1.1　低信噪比光谱重构

主成分分析(principal component analysis，PCA)是一种利用正交变换来将一组可能相关的变量转变为一组线性无关的变量(称之为主成分)的数学过程。主成分的数目要小于或等于原始变量的数目。基于 PCA 方法，本文按照下面方法来重构低信噪比恒星光谱：

选取文献[9]中的所有 O，B，A，F，G，K 和 M 型光谱，波长统一到 4 000～8 800 Å，步长为 1 Å(总的采样点数量 N =4 801)，得到插值后的流量 F。对 F 进行主成分分析，其中前 15 个特征向量的方差贡献率如表 1 所示。

表 1　前 15 个特征向量的方差贡献率

Table 1　The variance contribution rate of first 15 eigen vectors

序号	方差贡献率/%	累积方差贡献率/%
1	81.813 9	81.813 9
2	15.618 8	97.432 7
3	2.209 2	99.641 9
4	0.235 9	99.877 8
5	0.040 8	99.918 6
6	0.024 9	99.943 4
7	0.022 2	99.965 6
8	0.009 1	99.974 7
9	0.006 1	99.980 8
10	0.003 9	99.984 7
11	0.003 5	99.988 2
12	0.002 2	99.990 4
13	0.001 8	99.992 2
14	0.001 5	99.993 7
15	0.001 1	99.994 8

从表 1 中看出，这些特征向量的方差贡献率累积已超过 99.99%，方法后续试验采用这 15 个特征向量，其特征向量如图 1 所示。

对于任何一条恒星光谱，都可以上述步骤中获得到的特征向量进行重构。重构时将恒星光谱降维到特征空间内，然后分别乘以特征向量加和即可得到重构之后的光谱。为检验本文方法，分别选取了两条文献[9]中的两条模板光谱，通过叠加高斯噪声分别模拟信噪比 1，3，5，7，9 和 11 的光谱；图 2 及图 3 中六个子图中红色线为原始光谱、黄色线为叠加噪声之后的光谱、蓝色线为重构之后的光谱。从图中红线和蓝线的吻合程度可以看出该方法可以非常有效的对光谱尤其是低信噪比的恒星光谱进行重构。

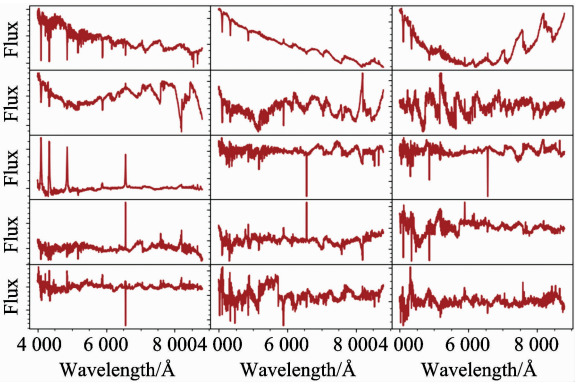

图 1　前 15 个特征向量

Fig. 1　The First 15 eigen-vectors

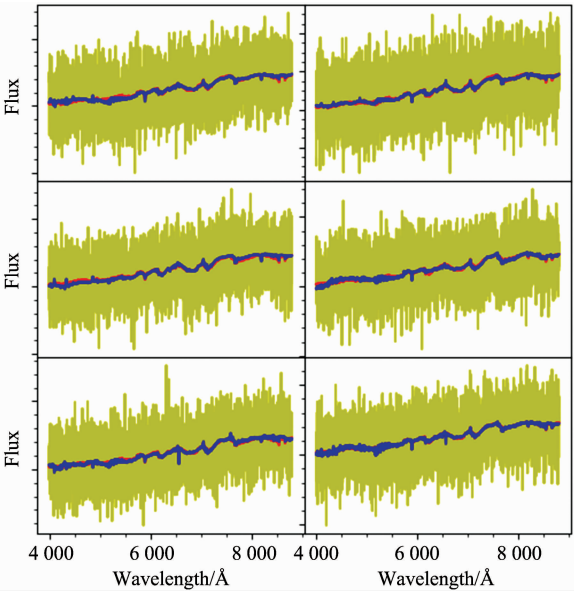

图 2　不同信噪比的 M 型光谱重构结果

Fig. 2　The result of spectral reconstruction for a M-type spectra with different SNRs

1.2　基于密度峰值的聚类方法

文献[10]中提出一种基于密度峰进行快速聚类的方法，该方法认为聚类中心应该具有以下两个特点：

(1)本身密度大，即它邻居的密度要小于该点；

(2)与其他密度更大的点距离相对更大。

对于样本集 S 中任何一个点 x_i，可以为其定义两个变量：ρ_i 及 σ_i。

局部密度：ρ_i。

本文采用的计算公式如下：

$$\rho_i = \sum_{j \in I_S \setminus \{i\}} X(d_{ij} - d_c)$$

其中函数

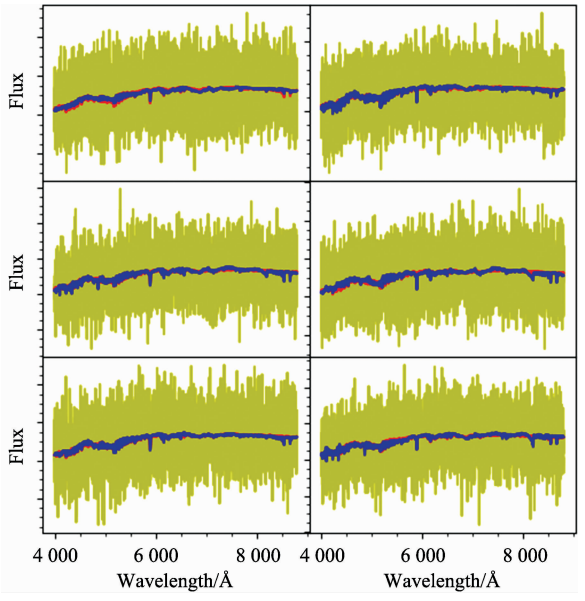

图 3 不同信噪比的 G 型光谱重构结果

Fig. 3 The result of spectral reconstruction for a G-type spectra with different SNRs

$$X(x)=\begin{cases}1 & x<0\\ 0 & x\geqslant 0\end{cases}$$

参数 d_c 定义为截断距离，从公式中可以看出 ρ_i 表示的是 S 中与 X_i 距离小于 d_c 的数据点的个数。

距离 σ_i：

设 $\{q_i\}_{i=1}^{n}$ 表示 $\{p_i\}_{i=1}^{n}$ 的一个降序排列下标序，即它满足

$$\rho_{q1}\geqslant\rho_{q2}\geqslant\cdots\geqslant\rho_{qN}$$

则可以定义

$$\delta_i=\begin{cases}\min\limits_{q_j,\ j<i}\{d_{q_i q_j}\} & i\geqslant 2\\ \max\limits_{j\geqslant 2}\{\delta_{qj}\} & i=1\end{cases}$$

这样，对于 S 中的每一个数据点 x_i，可为其计算(ρ_i，σ_i)，如图 4 所示例子，其中共包含了 28 个二维数据点如图 4(a)所示，将(ρ_i，σ_i)表示二维图(称之为决策图)中如图 4(b)所示。可以发现，1 号和 10 号点都具有较大的 ρ 和 δ 值，这两个点即为所有寻找的簇心，同时发现编号 26，27 和 28 三个数据点即为离群点，在图 4(b)中这三个点的 ρ 值很小同时 δ 值很大。

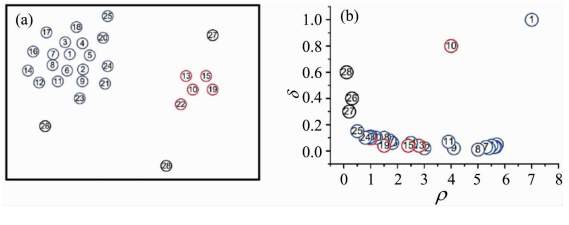

图 4 基于密度峰值的聚类例子

Fig. 4 A sample of clustering using the method of desnity peak

1.3 低信噪比巡天数据中特殊恒星光谱搜寻方法

提出的从低信噪比巡天数据中搜寻特殊恒星光谱的方法，综合前述恒星光谱重构以及光谱聚类方法，具体步骤如下：

(1) 对于所有低信噪比恒星光谱，首先利用前述方法进行光谱重构。

(2) 按照前面方法分别计算局部密度和距离，从决策图筛选离群数据，所得离群数据即为本文所要筛选的特殊恒星光谱。

2 实验检验

2.1 数据

本文数据同样来自文献[9]，随机选取其中编号范围 39～100(光谱型 FGK)的模板，随机添加噪声，构建 1 000 条光谱。然后在其中分别添加同信噪比的 136#(M8)，151#(Binary)，154#(CV)各 5 条。实验中，共有三组，每组 1 005 条光谱数据。该数据很好的模拟实际巡天样本中的数据分布。

2.2 结果

在上面构建的模拟数据中，应用本文提出的方法，其中光谱之间的距离采用余弦距离，截断距离 d_c 取 0.5，信噪比取 2，实验得到的决策图分别如图 5 所示。

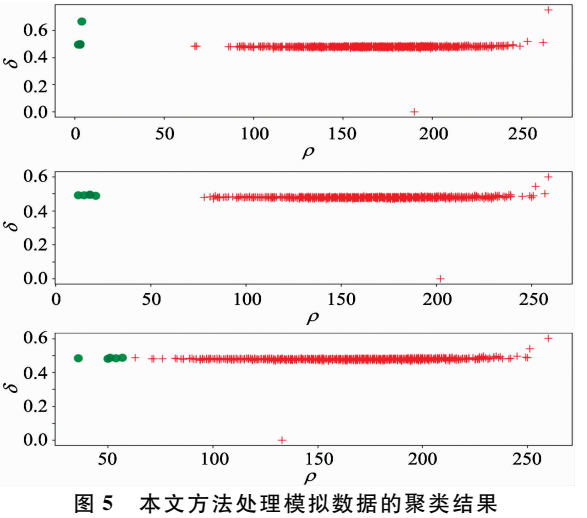

图 5 本文方法处理模拟数据的聚类结果

Fig. 5 The result of the method on mock data

从图 5 中可以看出，绿色的点(数量较少的特殊恒星光谱)在图中的分布明显偏移正常光谱的分布范围。也就是说，本文提出的方法在信噪比非常低(SNR＝2)的情况下，可以非常容易而且有效地发现其中数量较少的特殊恒星光谱。

3 结 论

分析了已有特殊恒星光谱搜寻方法的基础上，提出一种特别适用于低信噪比恒星光谱巡天数据中搜寻数量相对较少的特殊恒星的方法。该方法首先通过由高信噪比恒星光谱得到的特征向量进行光谱重构，提高光谱的信噪比；然后通过寻找密度峰的方法，间接找到其中的离群数据点。实验表明，本文方法在信噪比极低的光谱中仍然可以准确有效的搜寻出其中数量较少的特殊恒星光谱。本文方法也可应用于诸如 SDSS 和 Segue 等银河系巡天的光谱数据分析与挖掘中。

References

[1] Cui X Q, Zhao Y H, Chu Y Q, et al. Research in Astronomy & Astrophysics, 2012, 12(9): 1197.
[2] Luo A L, Zhao Y H, Zhao G, et al. Research in Astronomy & Astrophysics, 2015, 15(8): 1095.
[3] Wei P, Luo A L, Li Y B, et al. Monthly Notices of the Royal Astronomical Society, 2013, 431(2): 1800.
[4] Navarro S G, Corradi R L M, Mampaso A. Astronomy and Astrophysics, 2012, 538: A76.
[5] LIU Jie, PAN Jing-chang, WU Ming-lei, et al(刘 杰, 潘景昌, 吴明磊, 等). Spectroscopy and Spectral Analysis(光谱学与光谱分析), 2017, 37(12): 3904.
[6] Peng N, Zhang Y, Zhao Y, et al. Monthly Notices of the Royal Astronomical Society, 2012, 425(4): 2599.
[7] Shi J R, Luo A L, Li Y B, et al. Science China Physics, Mechanics and Astronomy, 2014, 57(1): 176.
[8] Mohamad A, Eva K G, et al. Astronomical Journal, 2014, 148(1): 8.
[9] Wei P, Luo A, Li Y, et al. Astronomical Journal, 2014, 147(5): 101.
[10] Rodriguez A, LAIO A. Science, 2014, 344(6191): 1492.

A Method to Search Special Stellar Spectra from Low Signal-to-Noise Ratio Spectral Sky Survey Data

WU Ming-lei[1], PAN Jing-chang[1*], YI Zhen-ping[1], WEI Peng[2]

1. School of Mechanical, Electrical & Information Engineering, Shandong University, Weihai, Weihai 264209, China
2. Key Laboratory of Optical Astronomy, NAOC, Chinese Academy of Sciences, Beijing 100012, China

Abstract Special stars are stars with anomalous metal abundance, the information of which is of great importance to the study of the origin of the universe, the evolution of the solar system and the evolution of life. Therefore, the search of special stars is an important goal in the large-scale survey project at home and abroad. Stellar spectra contain a wealth of information on the chemical composition, the physical property, and the movement state of stars, which is an important basis for conducting stellar studies. Stellar identification, classification, and the discovery of special stars are largely based on stellar spectral data. With the development of large-scale digital survey projects at home and abroad, such as LAMOST and SDSS, the data amount of stellar spectra has reached an unprecedented height. Such a large amount of data provide strong support for the discovery of special stars. Therefore, how to use these data to find the special, rare and even unknown types of stellar spectra rapidly and accurately is an important issue in astronomical research. Data mining is a technology that combines the pattern recognition, machine learning, statistical analysis and background knowledge of relevant experts to extract the potential unknown valuable information in the past. It has a natural advantage in dealing with big data. More and more data mining methods are applied to the survey data processing and analysis. At present, the data mining algorithms for special stars search mainly include stochastic forest, cluster analysis and outlier detection and so on. However, as the depth of the survey is expanded, the target of observation is getting darker and the signal-to-noise ratio of the observed spectrum accordingly lowers. There is a lot of useless information in the low signal-to-noise ratio spectrum, and the results obtained by directly analyzing and processing the relevant algorithms often have great deviations. Therefore, how to efficiently search out the special stellar spectra from a large number of low-SNR stellar data is an important issue nowadays. Due to the characteristics of the low-SNR stellar spectra themselves, a few studies are being done to search for the special stellar spectra. In order to solve this problem, a method based on principal component analysis (PCA) and the density peak approach is proposed to search special stellar spectra in low-S/N stellar data on the basis of careful study of the relevant methods. In this method, firstly, various types of high-SNR star spectra of O, B, A, F, G, K and M are selected, and then characteristic spectra are obtained by principal component analysis after wavelength unification and flux interpolation; secondly, the stellar spectra are reconstructed to obtain high-SNR spectra by using the first few characteristic spectra; finally, high-SNR spectra are clustered, and the outlier data is the special stellar spectrum. When clustering, this method uses a clustering method based on density peak for clustering and outlier mining with taking into account the characteristics of stellar spectral data itself. Experiments show that the proposed method can accurately search for a relatively smaller number of special stars in the low-SNR stellar data. At the same time, the proposed method can be applied to the spectral data analysis and mining of various galactic survey such as LAMOST and SDSS.

Keywords Galaxy survey; Outlier data mining; Low SNR spectra

* Corresponding author

(Received Jan. 22, 2018; accepted May 18, 2018)