第62卷 第2期 天 文 学 报 Vol.62 No.2
2021年3月 ACTA ASTRONOMICA SINICA Mar., 2021

doi: 10.15940/j.cnki.0001-5245.2021.02.004

基于NPSVM模型的超新星识别方法*

王精东[1] 陈星星[2] 梁吴颖[3] 黄泽锋[1] 全少武[1] 沈锦泫[1]
石雄辉[1†]

(1 广东海洋大学数学与计算机学院 湛江 524088)
(2 北京师范大学数学科学学院 北京 100875)
(3 湖南科技大学材料科学与工程学院 湘潭 411201)

摘要 巡天观测与高能物理、黑洞天文等领域均有密切的联系. 基于星系-超新星二分类问题, 研究光谱数据预处理, 结合余弦相似度改善PCA (Principal Component Analysis)光谱分解特征提取方法, 用SDSS (the Sloan Digital Sky Survey)、WISeREP (the Weizmann Interactive Supernova data REPository)组成的5620条光谱数据集训练支持向量机, 可以得到0.498%泛化误差的识别模型和新样本分类概率. 使用Neyman-Pearson决策方法建立NPSVM (Neyman-Pearson Support Vector Machine)模型可进一步降低超新星的漏判率.

关键词 超新星: 普通, 星系: 普通, 技术: 光谱学, 方法: 数据分析
中图分类号: P152; **文献标识码**: A

1 引言

遥远星系内的恒星在一般情况下亮度较低, 无法观测, 而其爆发成为超新星时则突然进入了可观测的范围, 此现象作为恒星演化的一个重要过程, 对研究恒星与高能天体物理具有重要影响[1]. 由于超新星的出现在时间与空间上均具有随机性, 使得超新星观测大多采用大规模巡天以获得大量对象的样本数据进行分析筛选.

超新星搜寻主要有可见光图像搜寻、测光、光谱认证3种方法[1–2], 其中, 光谱数据相对可见光图像, 具有更好的准确性, 且不用像可见光图像、测光一样需获取多日的数据进行比对校验, 是超新星搜寻的重要手段.

目前在运行的大型超新星巡天项目中, 国外有ZTF (the Zwicky Transient Facility)、Pan-STARRS (Panoramic Survey Telescope And Rapid Response System)、Gaia (Global Astrometric Interferometer for Astrophysics)等[3–5], 国内有TNTS (the THU-NAOC Transient Survey)、PTSS (the PMO-Tsinghua Supernova Survey)、AST3 (Antarctic Survey Telescope)等项目[6–8]. 这些大型巡天计划往往产生超大规模的数据, 现代天文观测对自动化分析的需求日益明显.

2020-06-05收到原稿, 2020-11-10收到修改稿
*广东省大学生创新创业训练计划项目(S201910566087)资助
†sxh70ww@163.com

光谱数据往往由于维数较多, 从而需要降维以方便计算, 常用的方法有主成分分析(Principal Component Analysis, PCA), 比如, 刘真祥等人直接采用PCA降维[9]; 屠良平等人基于PCA实现光谱分解并计算重构系数作为新特征[10]. 在识别方面, 可以采用深度信念网络, 刘真祥等人采用该方法在激变变星光谱分类问题上获得了97.4%的正确率[9]; 屠良平等人则采用基于局部孤立性因子(Local Outlier Factor, LOF)的离群点检测算法将超新星候选数降低到了实验总数的1%, 该实验每次随机将1条超新星光谱混入到5054条星系光谱中, 发现每次实验中超新星的LOF都排在实验样本集合的前1%[11]. 此外还可以使用支持向量机(Support Vector Machine, SVM)或人工神经网络(Artificial Neural Network, ANN)等方法[12–13].

经研究发现, 基于光谱分解的特征提取方法存在冗余特征的问题, 并提出了解决方案: 首先, 用光谱分解方法提取特征进行初筛, 然后计算所得特征系数在两个类别上的相似度, 并做进一步筛选; SVM在星系-超新星二分类问题中得到了99.502%的识别正确率. 考虑进一步降低漏判率, 结合Neyman-Pearson决策, 本文提出了NPSVM (Neyman-Pearson SVM)分类模型, 以适应不同条件与场景.

2 数据收集

超新星往往是从星系对象中筛选出来. 所以只需要收集超新星与星系两类的数据进行分析即可.

我们从WISeREP (the Weizmann Interactive Supernova data REPository)[1]上收集了SUSPECT (SUpernova SPECTrum), SNfactory (the Nearby Supernova Factory), SDSS-SNe (the Sloan Digital Sky Survey-II Supernova Survey)观测数据光谱[14], 使用了CfA (the Center for Astrophysics)[2]超新星计划[15]的一部分与Nugent等人构造的超新星模板[3][16]. 本文与文献[10–11]中仅采用Ia型超新星模板不同, 希望能够将模型推广到其他类型的超新星光谱识别上, 所以我们也应用了Nugent等人的其他类型超新星模板, 由于这些模板是由超新星同一时间的多个观测来源的光谱取平均得到, 并在时间上进行了插值[16], 更具有代表性, 因此将用于构造超新星的PCA模板, 而剩下的超新星实测光谱将用于分类模型的训练和测试.

星系数据上, 我们使用了SDSS-dr16得到的星系观测数据[17]. 这些星系光谱将既用于PCA模板库的构造又作为实验数据集的星系部分.

为了将数据统一在一个区间上进行分析, 本文筛选出静止波长范围在3809–7000 Å的光谱, 并获得了3427条超新星实测光谱和2193条星系光谱.

表1、2分别给出了超新星和星系实测光谱的类别数量和比例, 未给出子类别的记为untyped, 其中AGN (Active Galactic Nucleus)为活动星系核. 图1给出了超新星实测光谱相对光极大和数据集信噪比的主要分布, 图例中SNe表示超新星, Gal表示星系.

[1] https://wiserep.weizmann.ac.il
[2] https://www.cfa.harvard.edu/supernova/cfaspec_snIa_20120322.tar.gz
[3] https://c3.lbl.gov/nugent/nugent_templates.html

表 1 超新星(SN)类型统计
Table 1 Statistics of supernova types

Type	Amount	Proportion	Type	Amount	Proportion
SN I	3	0.09%	SN Ib/c	14	0.41%
SN Ia	3008	87.77%	SN II	41	1.20%
SN Ia-pec	46	1.34%	SN IIb	41	1.20%
SN Ia-91bg-like	26	0.76%	SN IIn	44	1.28%
SN Ia-91T-like	10	0.29%	SN IIP	44	1.28%
SN Ib	23	0.67%	SN IIL	2	0.06%
SN Ibn	3	0.09%	SN II-pec	41	1.20%
SN Ic	81	2.36%			

表 2 星系类型统计
Table 2 Statistics of galaxy types

Type	Amount	Proportion
Untyped	1305	59.48%
Starforming	614	27.99%
Starburst	179	8.16%
AGN	52	2.37%
Broadline	35	1.60%
AGN broadline	7	0.32%
Starforming broadline	2	0.09%

3 数据预处理

超新星形成机制包括白矮星吸积伴星产生的热核爆炸、大质量恒星的核塌缩爆炸, 两者均多发生在星系内. 所以超新星的光谱$\boldsymbol{S}$往往会与宿主星系的光谱$\boldsymbol{G}$发生叠加, 同时, 伴随有测量、大气条件、天光背景等引起的随机噪声$\boldsymbol{\varepsilon}$, 于是观测得到的光谱$\boldsymbol{M}$可以写成如下形式:

$$\boldsymbol{M} = \boldsymbol{S} + \boldsymbol{G} + \boldsymbol{\varepsilon}\,.$$

光谱作为对遥远天体的观测手段之一, 除了噪声影响外, 主要还存在以下影响因素:
(1)遥远天体的光谱红移;
(2)天体距离、超新星本身属性、所处演化阶段等对谱线强度的影响;
(3)不同望远镜数据处理方法的差异.

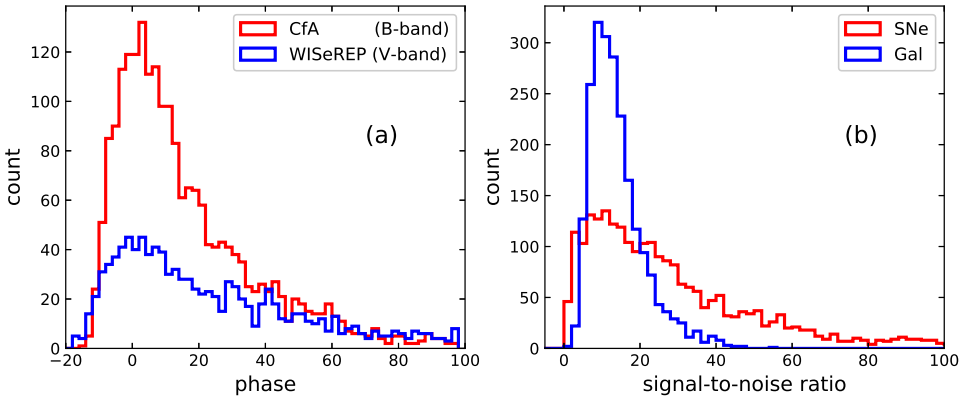

图 1 超新星相对光极大(a)和数据集信噪比(b)的分布直方图. (a)中收集自CfA的光谱(红色曲线)给出的是相对B波段光极大的天数, 收集自WISeREP的光谱(蓝色曲线)给出的是相对于V波段光极大的天数. (b)中红色曲线为超新星光谱的信噪比分布, 蓝色曲线为星系光谱的信噪比分布.

Fig. 1 Histogram of supernova phases (a) and signal-to-noise ratio of dataset (b). In panel (a), the spectra collected from CfA give the number of days relative to B-band maximum (red curve), while the spectra collected from WISeREP give the number of days relative to V-band maximum (blue curve). In panel (b), the red curve is the signal-to-noise ratio distribution of supernova spectra, and the blue curve is the signal-to-noise ratio distribution of galactic spectra.

因此本文对观测数据进行了以下预处理以减小上述因素对算法效果的影响:

(1)退红移: 对于红移为z、观测波长为$\lambda_{\rm obs}$的天体光谱, 计算其静止波长:

$$\lambda = \frac{\lambda_{\rm obs}}{1+z}\,;$$

(2)截断与插值: 将所有样本截断至静止波长覆盖范围的交集3809–7000 Å并进行线性插值;

(3)去除强窄线: 由于超新星的光谱是没有窄线的[10], 为了减少窄线的干扰, 本文基于3σ原则(将到均值的距离大于3倍标准差的点视为异常点)来检测窄线, 对窄线进行过滤并替换成均值, 循环多次以保证效果;

(4)去噪: 使用小波变换方法进行去噪;

(5)标准化: 对于均值为μ_i、标准差为σ_i的第i条观测数据$\boldsymbol{M}_{{\rm obs},i}$, 标准化后记为向量$\boldsymbol{M}_i = (\boldsymbol{M}_{i1}, \boldsymbol{M}_{i2}, \cdots, \boldsymbol{M}_{iq})$. 其第$j$个分量:

$$\boldsymbol{M}_{ij} = \frac{\boldsymbol{M}_{{\rm obs},ij} - \mu_i}{\sigma_i}\,, \quad i = 1,2,\cdots,N\,, \quad j = 1,2,\cdots,q\,,$$

其中, $\boldsymbol{M}_{{\rm obs},ij}$为观测数据向量$\boldsymbol{M}_{{\rm obs},i}$的第$j$个分量, N为观测样本的数量, q为光谱数据向量的维数. 图2展示了数据预处理的效果.

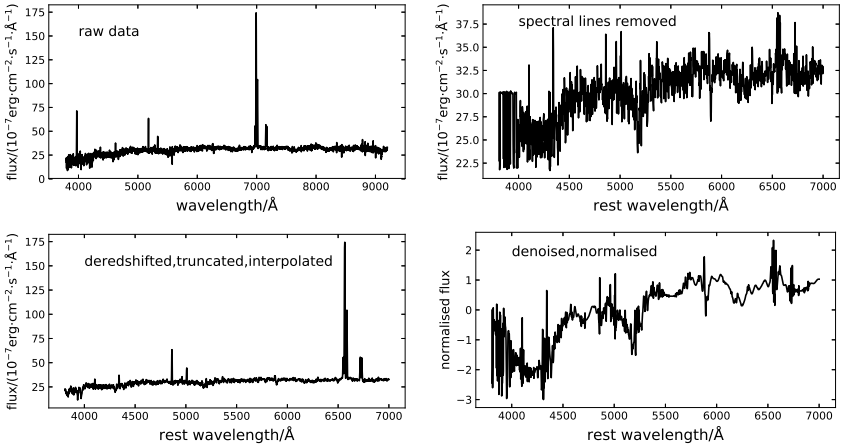

图 2 在数据预处理中, 一条光谱的原始数据将被退红移、截断、插值、去谱线、去噪、标准化.

Fig. 2 In the data preprocessing, the raw data of a spectrum will be deredshifted, truncated, interpolated, spectral lines removed, denoised and normalised.

4 特征提取

以下设行向量

$$\boldsymbol{M}_i = \boldsymbol{G}_i + \boldsymbol{S}_i + \boldsymbol{\varepsilon}_i = (\boldsymbol{M}_{i1}, \boldsymbol{M}_{i2}, \cdots, \boldsymbol{M}_{iq})$$

为第i条经过预处理的混合光谱, 其中, $\boldsymbol{G}_i$与$\boldsymbol{S}_i$分别表示组成该混合光谱的星系光谱和超新星光谱, $\boldsymbol{\varepsilon}_i$表示观测过程中的随机噪声.

4.1 光谱分解

PCA旨在通过线性变换得到新的变量, 即主成分, 并对主成分按照方差贡献率进行排序. 这些主成分满足: 各个主成分Y_j的方差达到最大且主成分之间协方差为0, 即第j个主成分Y_j的方差$\mathrm{Var}(Y_j) = \max$, Y_j与$Y_{j'}$的协方差$\mathrm{Cov}(Y_j, Y_{j'}) = 0$, $j > 1$, $j' = 1, 2, \ldots, j-1$. 对于第i条光谱$\boldsymbol{M}_i$, 其第j个主成分$Y_j = \boldsymbol{M}_i \boldsymbol{l}_j^{\mathrm{T}}$, 其中, $\boldsymbol{l}_j$称为主成分方向.

对超新星与星系模板库分别运用PCA, 则可以得到星系模板库的前m个主成分方向和超新星模板库的前n个主成分方向分别为$\{\boldsymbol{g}_u\}, \{\boldsymbol{s}_v\}, u = 1, 2, \cdots, m, v = 1, 2, \cdots, n$. 本文将主成分方向称为光谱特征.

其中, 由于Nugent先生提供的模板在不同类别的超新星之间, 光谱数量存在较大差异. 由于其中一类最多只有12条光谱数据, 于是对每一类在时间上等距选取12条光谱数据以用于构建超新星的PCA模板库.

图3自上而下展示了星系与超新星模板库按方差贡献率排序后的前8个光谱特征. 此时选取超新星光谱特征累计贡献率为95%的前8个光谱特征, 而对于星系成分选取累计贡献率为90%的前180个光谱特征. 此时, 可以将实测光谱看作是各光谱特征线性组合的近似[10,18], 即

$$\boldsymbol{M}_i \approx \boldsymbol{G}_i + \boldsymbol{S}_i = \sum_{u=1}^{m} a_{u,i} \boldsymbol{g}_u + \sum_{v=1}^{n} a_{m+v,i} \boldsymbol{s}_v \ . \tag{1}$$

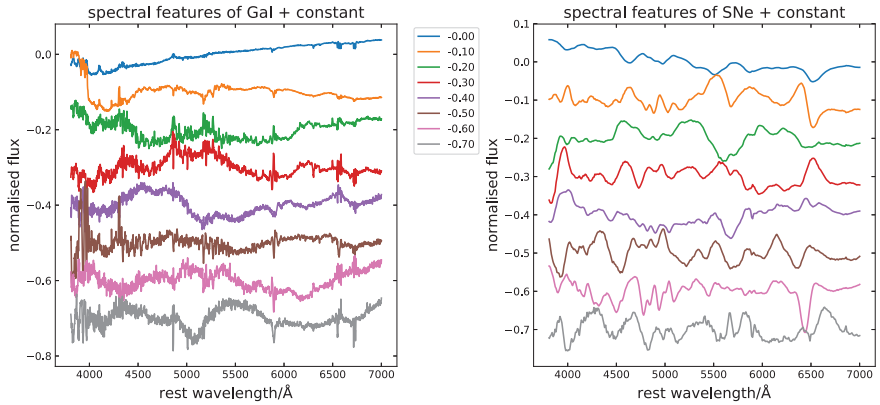

图 3　光谱特征. 自上向下为星系(左)与超新星(右)模板库按方差贡献率排序后的前8个光谱特征, 为了方便展示, 对每条光谱特征进行了偏移, 图例数字表示偏移量.

Fig. 3　Spectral feature. The galaxy (left) and supernova (right) templates are sorted by variance contribution and the first eight spectral feature are top down displayed in figure. In order to display conveniently, each spectral feature has an offset, and the numbers in legend represent the value of offset.

设第k个混合光谱的系数向量$\boldsymbol{a}_i=(a_{1,i},a_{2,i},\cdots,a_{m+n,i})^{\mathrm{T}}$, 记

$$\begin{cases}\boldsymbol{A}_{\mathrm{c}} &=(\boldsymbol{a}_1,\boldsymbol{a}_2,\ldots,\boldsymbol{a}_N)\,,\\ \boldsymbol{A} &=(\boldsymbol{g}_1^{\mathrm{T}},\boldsymbol{g}_2^{\mathrm{T}},\cdots,\boldsymbol{g}_m^{\mathrm{T}},\boldsymbol{s}_1^{\mathrm{T}},\boldsymbol{s}_2^{\mathrm{T}},\cdots,\boldsymbol{s}_n^{\mathrm{T}})\,,\\ \boldsymbol{M} &=(\boldsymbol{M}_1,\boldsymbol{M}_2,\cdots,\boldsymbol{M}_N)^{\mathrm{T}}\,,\end{cases}$$

由(1)式有$\boldsymbol{A}\cdot\boldsymbol{A}_{\mathrm{c}}=\boldsymbol{M}^{\mathrm{T}}$. 因此求解$\boldsymbol{A}_{\mathrm{c}}$即可求得各系数向量, 并实现光谱分解, 即计算

$$\boldsymbol{A}_{\mathrm{c}}=\boldsymbol{A}^{-1}\boldsymbol{M}^{\mathrm{T}}\,,$$

当$\boldsymbol{A}$不满足正交性时, 则可以用如下方法计算得到[10]:

$$\begin{cases}\boldsymbol{B} &=\boldsymbol{A}^{-1}(\boldsymbol{\beta}_1,\boldsymbol{\beta}_2,\cdots,\boldsymbol{\beta}_m,\boldsymbol{\beta}_{m+1},\cdots,\boldsymbol{\beta}_{m+n})\,,\\ \boldsymbol{A}_{\mathrm{c}} &=\boldsymbol{B}(\boldsymbol{\beta}_1,\boldsymbol{\beta}_2,\cdots,\boldsymbol{\beta}_m,\boldsymbol{\beta}_{m+1},..,\boldsymbol{\beta}_{m+n})^{\mathrm{T}}\boldsymbol{M}^{\mathrm{T}}\,,\end{cases}\tag{2}$$

其中, $(\boldsymbol{\beta}_1,\boldsymbol{\beta}_2,\cdots,\boldsymbol{\beta}_m,\boldsymbol{\beta}_{m+1},\cdots,\boldsymbol{\beta}_{m+n})$为$\boldsymbol{A}$经施密特正交化后所得的单位正交向量组.

通过光谱分解, 我们获得了将一条光谱数据映射到以$\boldsymbol{a}_i$为坐标向量的新特征空间的方法, 图4分别展示了星系与超新星的分解效果. 图5展示了部分变量组合的3维分布图.

4.2　特征分布的相似性度量

假设已知两条连续的概率密度曲线$p_1(x)$、$p_2(x)$, 对p_1、p_2以Δx为步长等距离离散可分别得到向量$\vec{p}_1$、$\vec{p}_2$. 当$\Delta x\to 0$时, 他们能够很好地表征出该曲线的特征. 故定义两条概率密度曲线的相似度:

$$\mathrm{Sim}(p_1,p_2)=\cos<\vec{p}_1,\vec{p}_2>=\frac{\vec{p}_1\cdot\vec{p}_2}{|\vec{p}_1||\vec{p}_2|}\,.\tag{3}$$

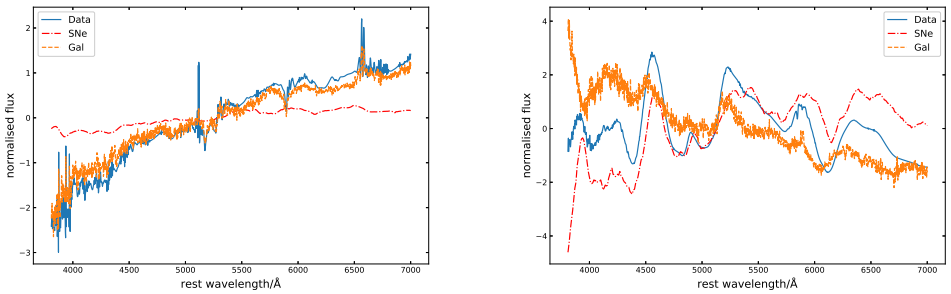

图 4 星系光谱(左)与超新星光谱(右)分解效果, 图中蓝色曲线为经过数据预处理的光谱, 红色曲线是分解出来的超新星成分, 橙色曲线是分解出来的星系成分.

Fig. 4 Decomposition of Gal spectrum (left) and SN spectrum (right). The blue curve is the spectrum after data preprocessing, the red curve is the supernova component, and the orange curve is the galaxy component of the decomposition.

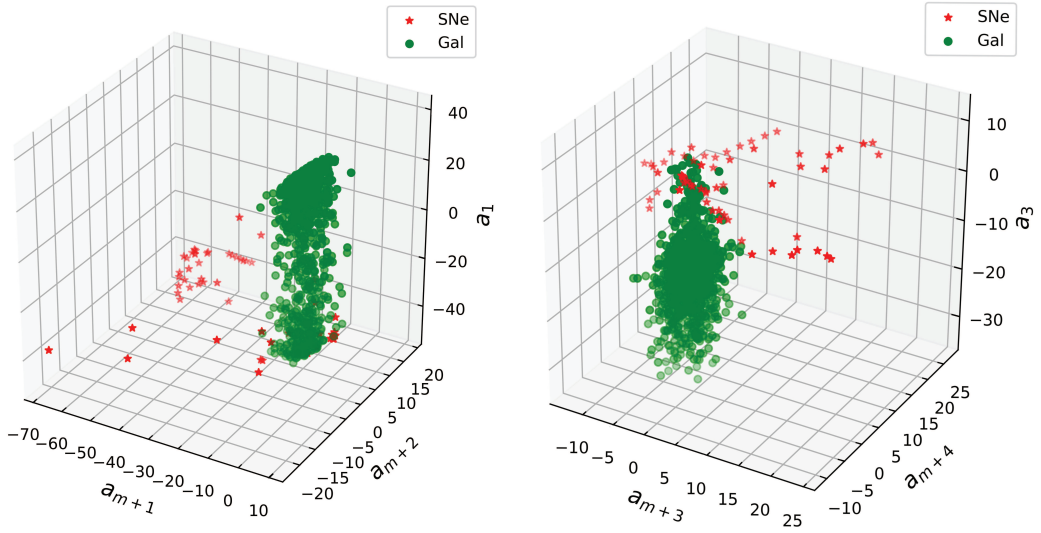

图 5 部分变量组合的3维分布图

Fig. 5 Three dimensional distributions of partial variables combination

图6为部分特征系数的频率分布直方图, 从图中可以看到部分特征系数出现冗余的现象, 这些特征系数的分布在两个类别之间的区别并不明显, 甚至接近重合. 这些特征系数的存在对分类任务并没有益处甚至会对分类效果造成一定的干扰.

于是需要度量一个特征系数的分布在两个类别之间差异的不明显程度, 即特征系数的分布相似程度, 以进一步筛选特征系数. 而(3)式恰好可以满足这一需求.

图7为相似度分布直方图. 于是, 只要设定阈值s_0, 保留Sim $\leqslant s_0$的特征系数, 即可过滤低效特征.

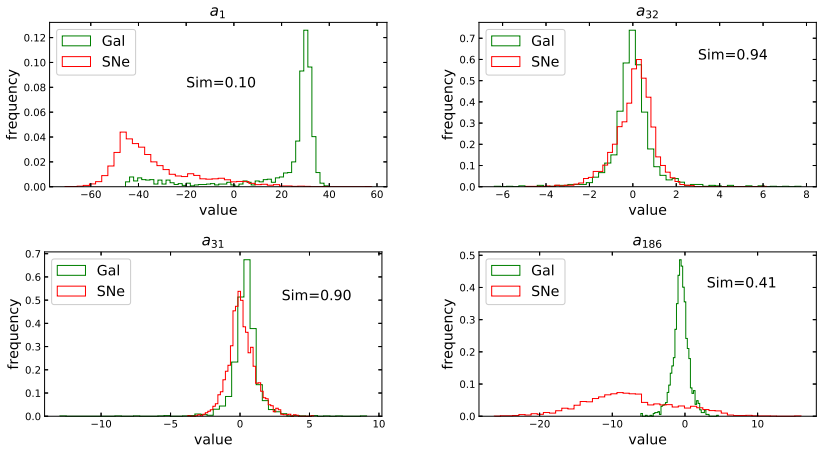

图 6　部分特征系数的频率分布直方图

Fig. 6　Frequency distribution histogram of partial feature coefficients

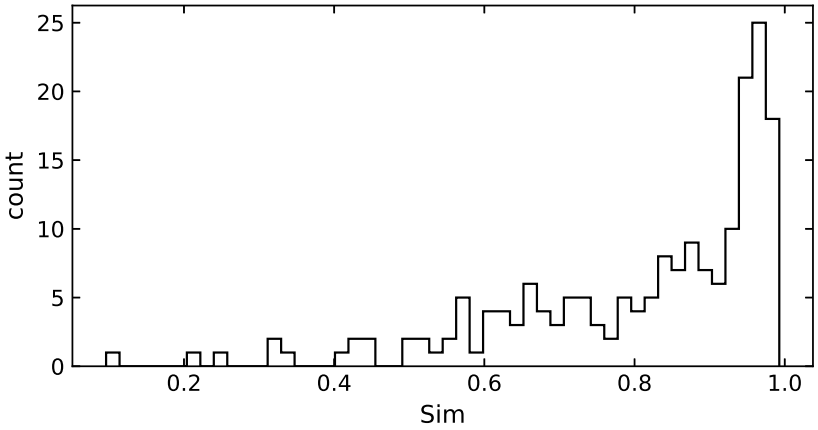

图 7　相似度分布直方图

Fig. 7　Histogram of similarity

5　光谱识别

由图5可见, 在光谱分解得到的特征空间下, 光谱识别问题可近似为一个线性可分的分类模型. 对此, 本文采用了线性支持向量机进行分类训练.

5.1　支持向量机

线性可分是指在特征空间中能够找到一个由下式表示的超平面来正确地描述分类任务:

$$\boldsymbol{w}\cdot\boldsymbol{x}^{\mathrm{T}}+b=0\,, \tag{4}$$

其中, $\boldsymbol{w}$为超平面的法向量, $\boldsymbol{x}$为坐标向量. 在光谱分类问题中, 用第i条预处理光谱$\boldsymbol{M}_i$提取的特征$\boldsymbol{a}_i$来代表该待分类光谱, 即有$\boldsymbol{x}=\boldsymbol{a}_i$.

于是一个样本点$\boldsymbol{x}$到超平面的距离为:

$$r = \frac{\boldsymbol{w}\cdot\boldsymbol{x}^{\mathrm{T}}}{||\boldsymbol{w}||}\,. \tag{5}$$

若该超平面能够正确分类, 即对于任一样本$(\boldsymbol{x}_i, y_i)$, $y_i \in \{-1, 1\}$, 当$\boldsymbol{x}_i$为星系光谱时, $y_i = -1$, $\boldsymbol{w}\cdot\boldsymbol{x}_i^{\mathrm{T}} + b < 0$; 当$\boldsymbol{x}_i$为超新星光谱时$y_i = 1$, $\boldsymbol{w}\cdot\boldsymbol{x}_i^{\mathrm{T}} + b > 0$.

只需要一定的事先变换, 样本集总可以满足:

$$\begin{cases} \boldsymbol{w}\cdot\boldsymbol{x}_i^{\mathrm{T}} + b \geqslant 1,\ y_i = 1. \\ \boldsymbol{w}\cdot\boldsymbol{x}_i^{\mathrm{T}} + b \leqslant -1,\ y_i = -1. \end{cases}$$

此时, 根据(5)式, 由该超平面所确定的分类间隔为$2/||\boldsymbol{w}||$. 对于分类问题, 我们希望这个分类间距尽可能大, 而这等效于最小化$\frac{1}{2}||\boldsymbol{w}||^2$. 因此我们只需要求解:

$$\begin{cases} \min_{\boldsymbol{w},b} \quad \frac{1}{2}||\boldsymbol{w}||^2\,, \\ \text{s.t.} \quad\ y_i(\boldsymbol{w}\cdot\boldsymbol{x}_i^{\mathrm{T}} + b) \geqslant 1\,, \end{cases} \tag{6}$$

引入拉格朗日乘子η_i, 由拉格朗日乘子法有对偶问题

$$\begin{cases} \max \quad \sum_{i=1}^{N}\eta_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{i'=1}^{N}\eta_i\eta_{i'}y_iy_{i'}\boldsymbol{x}_i\cdot\boldsymbol{x}_{i'}^{\mathrm{T}}\,, \\ \text{s.t.} \quad \sum_{i=1}^{N}\eta_i y_i = 0,\ \eta_i \geqslant 0,\ i' = 1, 2, \cdots, N\,. \end{cases} \tag{7}$$

考虑防止过拟合或不是完全线性可分的情况下, 引入"松弛变量"$\xi_i \geqslant 0$, 求解如下规划问题:

$$\begin{cases} \min_{\boldsymbol{w},b,\xi} \quad \frac{1}{2}||\boldsymbol{w}||^2 + C\sum_{i=1}^{N}\xi_i\,, \\ \text{s.t.} \quad\ y_i(\boldsymbol{w}\cdot\boldsymbol{x}_i^{\mathrm{T}} + b) \geqslant 1 - \xi_i, \quad \xi_i \geqslant 0\,, C > 0\,, \end{cases} \tag{8}$$

即"软间隔支持向量机"[12], 其中, C为常数, 表示约束程度, C越大时松弛变量要越小, 通常取$C = 1$.

在上文的特征提取方法下, 选取数据集的75%作为训练集, 剩余作为测试集. 迭代选取不同s_0进行SVM的训练, 得到了如图8的误判率变化曲线, 并在表3中展示了3个模型A、B、C的详细效果.

表 3 SVM在测试集上的表现

Table 3 The SVM performance on the test set

Model	s_0	Features	Mistake	Precision	Recall
A	0.895	103	0.641%	99.650%	99.188%
B	0.980	174	0.498%	99.422%	99.768%
C	1.000	188	0.783%	99.419%	99.304%

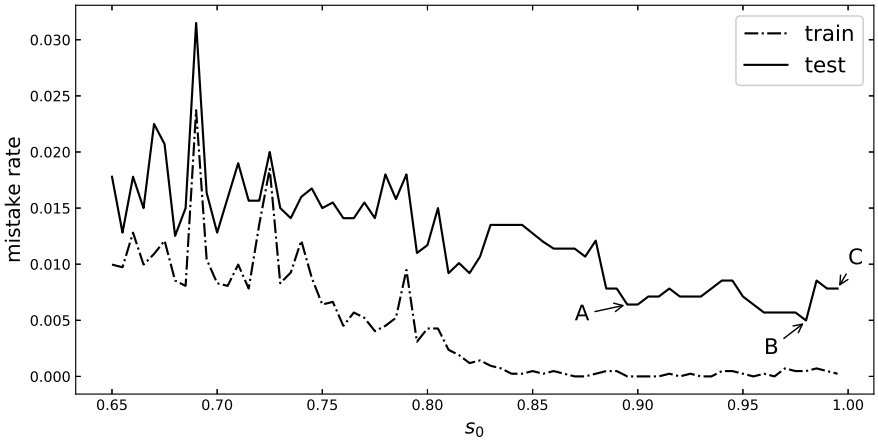

图 8 误判率曲线, s_0为相似度阈值

Fig. 8 Mistake rate curve, s_0 is the threshold of the similarity Sim

显然, 相似度过滤算法能够进一步提高降维结果的质量, 当s_0选取适当时, 仍能保证分类效果而数据维数较少. 在实际情况中, 可以视条件选择合适的模型. 比如, 当设备计算资源有限时采用算法空间复杂度较低的A模型, 当计算资源足够且希望漏判率尽可能小则使用漏判率最小的B模型.

由(5)式可以得到对于样本$\boldsymbol{x}$到决策面的距离$r(\boldsymbol{x})$. 图9展示了采用B模型时所得距离分布密度图. SVM可以视为一个以原点为决策点的分类模型, 在直线L_1右侧即判断为正, 在L_1左侧判断为负.

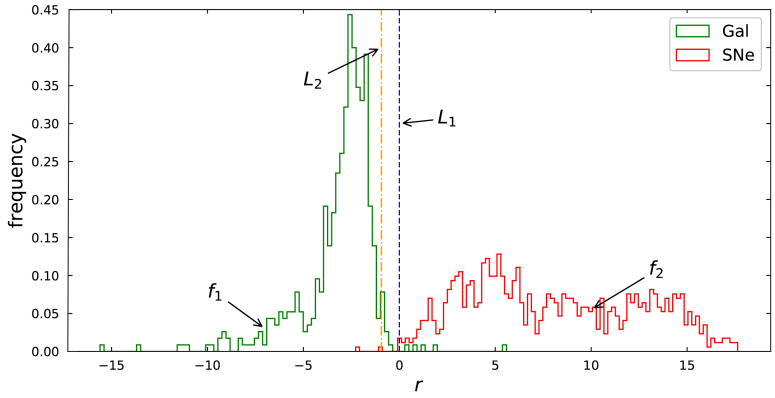

图 9 采用B模型时, 样本点到SVM超平面的距离r的分布. f_1为星系样本集上的距离分布曲线, f_2为超新星样本集上的距离分布曲线, L_1和L_2分别为SVM、NPSVM的决策线.

Fig. 9 Distribution of distance r from the sample point to the SVM hyperplane when using the model B. f_1 is the distance distribution on the set of galaxies, f_2 is the distance distribution on the set of supernovae. L_1, L_2 are the decision-making lines of SVM and NPSVM, respectively.

当样本数足够大时, 决策距离r的频率分布曲线能够近似于概率密度曲线, 即$f_1 \to p_1(r)$、$f_2 \to p_2(r)$, 其中, f_1为星系样本集上r的频率分布曲线, f_2为超新星样本集上r的频率分布曲线; $p_1(r)$为当样本$\boldsymbol{x}$来自星系总体时, $r = r(\boldsymbol{x})$在实数域$\mathbb{R}$上的概率密度曲线, $p_2(r)$为$\boldsymbol{x}$来自超新星总体时r的概率密度曲线.

此时, 设$P_1[r(\boldsymbol{x})]$是样本$\boldsymbol{x}$来自星系总体的概率, $P_2[r(\boldsymbol{x})]$为样本$\boldsymbol{x}$来自超新星总体的概率. 概率$P_1(r), P_2(r)$分别定义如下:

$$P_1(r) = \frac{p_1(r)}{p_1(r) + p_2(r)}\,, \quad P_2(r) = \frac{p_2(r)}{p_1(r) + p_2(r)}\,, \tag{9}$$

且给定$\boldsymbol{x}$时, 有$P_1(r) + P_2(r) = 1$.

因此当样本量足够大时, 用$f_1(r)$近似$p_1(r)$, 用$f_2(r)$近似$p_2(r)$, 即可通过SVM模型计算一个样本属于两个类别的概率. 当有多个超新星候选体待观测时, 可以将候选体按照超新星的分类概率$P_1(r)$排序, 制定观测计划, 从而更高效地寻找超新星.

5.2 NPSVM分类模型

来自超新星样本集的样本如果被误判为星系, 即为超新星的漏判. 对于超新星巡天, 有时希望超新星的漏判率尽可能低. 为了进一步降低漏判率, 不妨将距离r看作是经过SVM映射所得的新特征.

设二元划分$R = (R_1, R_2)$, 当对于一个新的样本$(\boldsymbol{x}_i, y_i)$, $r(\boldsymbol{x}_i)$经过划分R后判断为星系时, 记作$r \in R_1$; 当$r(\boldsymbol{x}_i)$经过R后判断为超新星时, 记为$r \in R_2$.

当$\boldsymbol{x}_i$来自星系总体时, 经过划分R后判断为超新星则是误判, 假设此事件发生的概率为$P_{\mathrm{e}1}$; 同理, 对于超新星总体, 有$P_{\mathrm{e}2}$. 由概率论有:

$$P_{\mathrm{e}1} = \int_{R_2} p_1(r)\mathrm{d}r\,, \quad P_{\mathrm{e}2} = \int_{R_1} p_2(r)\mathrm{d}r\,. \tag{10}$$

为了进一步降低超新星漏判, 对于期望的超新星漏判上限ε_0, 希望有$P_{\mathrm{e}2} = \varepsilon_0$, 此时要寻找满足如下条件的最优划分$R$, 即Neyman-Pearson决策:

$$\begin{cases} \min & P_{\mathrm{e}1} = \int_{R_2} p_1(r)\mathrm{d}r\,, \\ \text{s.t.} & P_{\mathrm{e}2} = \varepsilon_0\,. \end{cases} \tag{11}$$

为了求解该问题, 可以构建拉格朗日函数$\gamma = P_{\mathrm{e}1} + \eta(P_{\mathrm{e}2} - \varepsilon_0)$, 又因为$1 - \int_{R_2} p_1(r)\mathrm{d}r = \int_{R_1} p_1(r)\mathrm{d}r$, 则有

$$\gamma = 1 - \eta\varepsilon_0 + \int_{R_1} [\eta p_2(r) - p_1(r)]\mathrm{d}r\,, \tag{12}$$

其中, η为引入的拉格朗日乘子.

由拉格朗日乘子法有:

$$\frac{\partial\gamma}{\partial r} = 0 \Rightarrow \eta = \frac{p_1(r_0)}{p_2(r_0)}\,, \quad \frac{\partial\gamma}{\partial\eta} = 0 \Rightarrow \int_{R_1} p_2(r)\mathrm{d}r = \varepsilon_0\,, \tag{13}$$

其中, r_0由$\int_{R_1} p_2(r)\mathrm{d}r = \int_{-\infty}^{r_0} p_2(r) = \varepsilon_0$解出.

于是, 由(11)式和(13)式有最优划分:

$$\begin{cases} \eta < p_1(r)/p_2(r)\,, & r \in R_1 \\ \eta > p_1(r)/p_2(r)\,, & r \in R_2 \end{cases}, \quad \begin{cases} R_1 = (-\infty, r_0) \\ R_2 = (r_0, +\infty) \end{cases}, \tag{14}$$

其中, $p_1(r)/p_2(r)$称为似然比. 于是, 通过调整ε_0即可适应不同的需求.

当样本数N足够大时, 有$f_1 \to p_1(r)$、$f_2 \to p_2(r)$. 另外, 在实际计算中, 由于采样的不足, 可能出现$f_1(r) = 0$和$f_2(r) = 0$的情况, 为了避免分母为0, 频率近似后可对决策函数(14)式做变换, 于是有:

$$\begin{cases} h(r) > 0\,, & y_i = 1 \\ h(r) < 0\,, & y_i = -1 \end{cases}, \tag{15}$$

其中, $h(r) = f_1(r_0)f_2(r) - f_1(r)f_2(r_0)$, r_0满足$\int_{-\infty}^{r_0} f_2(r)\mathrm{d}r = \varepsilon_0$, 本文称(15)式为NPSVM模型.

若令$\varepsilon_0 = 0.002$, 则得到图9中的直线L_2, 并计算$h(r)$得到其频数分布图如图10, 此时与SVM模型的比较效果如表4.

计算得到超新星此时的漏判率为0.116%. 显然, 效果比预期的$\varepsilon_0 = 0.2\%$要好很多. 但是, 星系被判为超新星的频率为0.920%, 而该值采用普通的SVM时, 仅有0.356%; 整体误判率也由0.498%上升到0.854%. 故而, 此方法应当根据实际情况中的效益和需要适当调整期望的漏判上限ε_0.

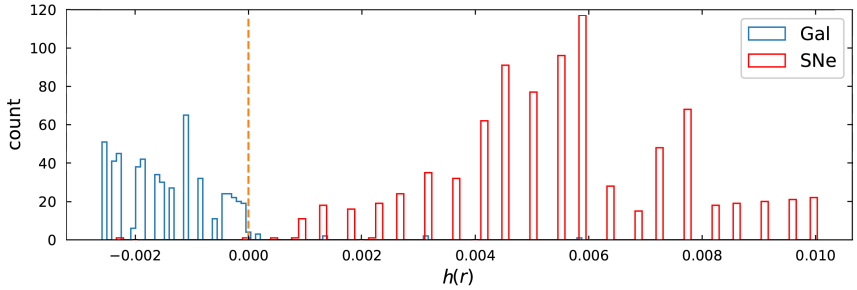

图 10　$h(r)$的分布直方图

Fig. 10　Histogram of $h(r)$

表 4　SVM与NPSVM效果比较

Table 4　Comparison between SVM and NPSVM

Model	Generalization error	Gal missing	SNe missing
SVM	0.498%	0.356%	0.232%
NPSVM	0.854%	0.920%	0.116%

6 总结与展望

本文通过数据预处理、特征提取、分类模型建立和优化, 发现了光谱分解方法的冗余特征并给出了基于余弦相似度的优化方案, 得到了一个效果良好的超新星识别模型. 相对于其他研究, 仅采用SVM模型时, 本文能够由决策距离分布给出一个样本的分类概率, 以方便设置超新星的后续观测优先级. 而追求较小超新星漏判率时, 本文提出的NPSVM模型也能够有效限制其漏判率.

但是在实践中, 虽然NPSVM模型可以降低漏判率, 但也会增加将星系判为超新星的概率. 超新星的最终证认往往还要通过人工证认或后续观测进行确认, 而星系被判为超新星会增加人力与时间成本, 因而需要根据实际情况进行漏判率的限制. 如果希望整体误判率最小, 还可以选择仅采用SVM模型而不采用Neyman-Pearson决策.

值得一提, 本文所涉及的PCA、SVM算法, 当数据量越大时, 在统计意义下越可靠, 效果也越好. 预期随着更多的超新星发现, 模型效果将不断提高, 从而形成正反馈机制推动超新星的发现.

致谢 感谢审稿人对文章提出的宝贵建议, 使得文章的质量有了显著的提高. 感谢WISeREP、CfA、SDSS公开的光谱数据和Nugent公开的超新星模板和相关工作者的付出.

参 考 文 献

[1] Branch D, Wheeler J C. Supernova Explosions. Berlin: Springer, 2017: 3-12
[2] Villar V A, Berger E, Miller G, et al. ApJ, 2019, 884: 83
[3] Bellm E C, Kulkarni S R, Graham M J, et al. PASP, 2019, 131: 018002
[4] Kaiser N, Aussel H, Burke B E, et al. Proceedings Volume 4836, Survey and Other Telescope Technologies and Discoveries. Hawaii: SPIE, 2002, 4836: 154
[5] Brown A G A, Vallenari A, Prusti T, et al. A&A, 2018, 616: A1
[6] Zhang T M, Wang X F, Chen J C, et al. RAA, 2015, 15: 215
[7] Zhang J J, Zhang K X, Lu J M, et al. ATel, 2019: 12810
[8] Liang E S, Zhang H, Yu Z Y, et al. AJ, 2020, 159: 201
[9] 刘真祥, 荣容, 许婷婷, 等. 云南民族大学学报(自然科学版), 2017, 26: 162
[10] 屠良平, 罗阿理, 吴福朝, 等. 中国科学: 物理学力学天文学, 2010, 40: 1282
[11] 屠良平, 罗阿理, 吴福朝, 等. 光谱学与光谱分析, 2009, 29: 3420
[12] Kou S H, Chen X Z, Liu X W. ApJ, 2020, 890: 177
[13] Liu C, Cui W Y, Zhang B, et al. RAA, 2015, 15: 1137
[14] Yaron O, Gal-Yam A. PASP, 2012, 124: 668
[15] Blondin S, Matheson T, Kirshner R P, et al. AJ, 2012, 143: 126
[16] Nugent P, Kim A, Perlmutter S. PASP, 2002, 114: 803
[17] Aguado D S, Ahumada R, Almeida A, et al. ApJS, 2019, 240: 23
[18] Pace Z J, Tremonti C, Chen Y M, et al. ApJ, 2019, 883: 82

A Supernova Recognition Method Based on NPSVM

WANG Jing-dong[1] CHEN Xing-xing[2] LIANG Wu-ying[3] HUANG Ze-feng[1]
QUAN Shao-wu[1] SHEN Jin-xuan[1] SHI Xiong-hui[1]

(1 College of Mathematics and Computer Science, Guangdong Ocean University, Zhanjiang 524088)
(2 School of Mathematical Sciences, Beijing Normal University, Beijing 100875)
(3 School of Materials Science and Engineering, Hunan University of Science and Technology, Xiangtan 411201)

Abstract Sky survey is closely related to the developments of many domains such as high energy physics and black hole astrophysics. In order to solve the classification problem between galaxy and supernova, an available supernova recognition method based on NPSVM (Neyman-Pearson Support Vector Machine) has been proposed. The dataset, which is collected from WISeREP (the Weizmann Interactive Supernova data REPository), SDSS (the Sloan Digital Sky Survey) and supernova templates made by Nugent, has 3427 supernova spectra and 2193 galaxy spectra. After preprocessing spectral data, the decomposed spectrum feature based on the Principal Component Analysis (PCA) is extracted, and the redundant features are decreased with the cosine similarity method. The classification model based on Support Vector Machine (SVM) has a low level of generalization error evaluated 0.498%, and can calculate the classification probability for a new sample. Furthermore, the improved NPSVM model can limit the missing rate on supernovae with the Neyman-Pearson criterion.

Key words supernovae: general, galaxies: general, techniques: spectroscopic, methods: data analysis