第 30 卷 第 1 期
2012 年 2 月

天 文 学 进 展
PROGRESS IN ASTRONOMY

Vol. 30, No. 1
Feb., 2012

文章编号：1000-8349(2012)01-094-105

光谱数据挖掘中的特征提取方法

李乡儒

(华南师范大学 数学科学学院，广州 510631)

摘要：特征提取是对光谱测量数据成分的分解、重组和选择的过程，它是光谱数据挖掘中的一个关键环节，不仅决定着后续处理的质量、效率、系统复杂度和稳健性，也关系到能够挖掘到什么知识和处理结果物理意义的可解释性。按照特征表达方式将已有方法分为 3 类：统计约简法，特征谱法和谱线法，并对这些方法的基本原理、适用性、优缺点及其在光谱数据挖掘中的应用作了综述和分析。另外，亦从方法的“时”、“频”分析能力方面探讨了不同方法的特点，例如，物理意义的易解释性、对波长定标畸变和流量定标畸变的敏感性等。

关 键 词：光谱；数据挖掘；特征提取

中图分类号：P11 **文献标识码**：A

1 引 言

随着传感器技术的快速发展和多个大型巡天计划的逐步实施，天文数据以“雪崩”之势增长 [1, 2]，由此导致了天文数据自动挖掘方法研究的必要性和迫切性。而已有数据挖掘结果对天文研究的促进作用则更进一步激发了人们对天文数据挖掘研究的热情 [1−4]。

特征提取是光谱数据挖掘中的一个核心环节 [5]，它对海量天体光谱数据处理的效率、准确性，以及分析方法对光谱中的噪声干扰、波长定标和流量定标不完备所导致的光谱畸变等因素的稳健性均有重要影响。“特征提取”包括特征的转换和选择两个环节，重在提取与分析目标有关的信息，尽可能剔除其它与当前任务无关的数据成分，并把信息转化为适合后续分析的表达方式，它直接关系到光谱挖掘结果的精确性/准确性和系统的复杂度 [6, 7]。例如，在光谱分类中，特征提取的质量不仅影响着最后的分类准确率，也决定着分类器的复杂性和效率。所以，本文探讨的光谱特征提取问题，是海量天文观测光谱的自动处理、信息提取、高效计算和共享等数据挖掘技术的关键。

在光谱数据挖掘中，特征提取包括 3 个关键成分：(1) 特征的检测和定位；(2) 特征的表

收稿日期：2011-05-10； **修回日期**：2011-11-07

资助项目：国家自然科学基金 (61075033)；广东省自然科学基金 (S2011010003348)；中国科学院模式识别国家重点实验室开放基金 (201001060)；华南师范大学教学改革项目 (2009jg28)

达；(3) 特征选择。虽然文献中有许多关于光谱特征提取的研究，但是按照特征的表达方式，本质上可以分为 3 类：统计约简法，特征谱法和谱线法。本文对上述各种方法及其在光谱数据处理中的应用作了综述，并对其优势、局限性和适用性作了分析。

2　统计约简法

这是应用最广泛的一类光谱特征提取方案，优点是易于操作和使用。该类方法本质上是对天体辐射能量进行分解、重组和取舍的过程，其目的是尽可能去除冗余、噪声，并将信号转化为利于后续处理的表达方式。常用的统计约简法有主成分分析、小波变换、流形学习以及有监督的相关向量机、支持向量机和判别分析法等。

2.1　主成分分析

在实际问题中，研究目标往往有多个测量指标，且不同指标之间有一定的相关性，这势必增加问题的复杂性。通过主成分分析 (Principal Component Analysis，简称 PCA)[8] 可将已有的众多指标进行分解、重组，形成一系列线性无关的综合指标，并按照它们反映原始信号所蕴含信息的能力从高到低进行排序。如果在数据分析中仅仅使用其中数个描述能力较强的合成指标，则达到了数据约减和特征提取的目的。并将这些合成指标依次称为第一主成分、第二主成分等。

这里的 “信息” 是指观测数据之间的差异、可区分性，例如，正是由于不同天体的观测光谱之间一般存在一定的差异，所以能够据此对观测目标的类别、红移等做出估计。假设 $X=\{x_i, i=1,\cdots,n\}$ 是一批观测数据，其采样总体的协方差矩阵是 Σ_P ，它反映了数据在观测空间中不同方向上的可区分性。但是，由于在应用中观测总体的真实协方差矩阵 Σ_P 往往是未知的，所以一般采用它的无偏估计进行 PCA 分析：

$$\hat{\Sigma}_P=\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})(x_i-\bar{x})^T, \tag{1}$$

其中 $x_i=(x_{i,1},x_{i,2},\cdots,x_{i,m})^T$，$x_{i,j}\in R$，$\bar{x}=\sum_{i=1}^{n}x_i/n$。$\hat{\Sigma}_P$ 是一个对称半正定矩阵，假设它的 m 个特征值和单位特征向量分别是 $\{(\lambda_i,v_i), i=1,2,\cdots,m\}$：

$$\hat{\Sigma}_P v_i=\lambda_i v_i, \tag{2}$$

则特征值 λ_i 反映了数据在 v_i 方向上的可区分性，λ_i 越大，数据在 v_i 方向上的区分性越强 (图 1 (a))。为了便于阐述，不妨假设

$$\lambda_1\geq\lambda_2\geq\cdots\geq\lambda_m. \tag{3}$$

需要注意的是，在式 (1) 中需要对每个观测数据减去均值 (该操作称为数据的中心化处理)，否则，分析的结果会对观测数据的整体平移变换敏感，而这显然不合理 (图 1(b))。

主成分分析在天体光谱数据挖掘中的应用包括两方面：数据压缩和特征谱构造。在数据

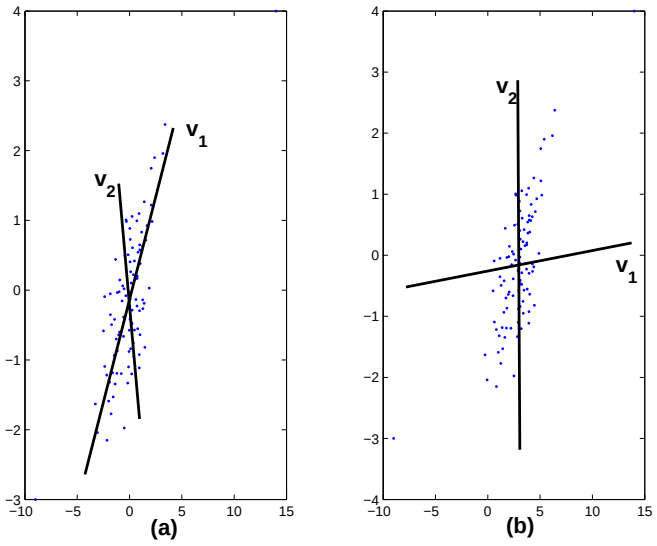

图 1 主成分分析效果分析：每个点代表一个观测数据，每条实线代表一个主成分，其方向是协方差矩阵的特征向量，长度与相应的特征值平方根成正比。(a) 对随机生成的一批数据进行 PCA 分析，由结果可见主方向和相应的特征值反映了数据的典型可区分方向和在相应方向上的可区分性。(b) 对图 (a) 中的观测数据沿横轴移动 3 个单位，且在 PCA 分析时不进行中心化处理，由结果可见，所得数据成分并不能正确反映数据在该方向上的可区分性。

压缩中，一般根据问题需要和某个选定的方差贡献率 $1 > \alpha > 0$，使

$$k = \min\{l : \sum_{i=1}^{l} \lambda_i / \sum_{i=1}^{m} \lambda_i \geq \alpha\}. \tag{4}$$

然后，对于任一观测数据 x，通过以

$$z = (x^T v_1, \cdots, x^T v_k)^T \tag{5}$$

代替 x 做统计分析达到数据压缩的目的，以利于高效计算，并抑制噪声等干扰因素对分析结果的不利影响。作为一种高效，易于使用的数据压缩方法，PCA 已经广泛地应用在光谱数据挖掘中，例如，恒星参数估计 [9–11]，恒星分类 [12]，星系光谱的分类 [13–15]，星系模型参数估计及其对尘埃和噪声的敏感性 [16]，吸收线和发射线光谱的分类 [17]，恒星形成历史 [18]，星系与类星体光谱的识别 [19]，类星体光谱 Lyα 线丛连续谱估计 [20–22]，类星体光谱的分类 [23]，以及低红移类星体的发射线特点 [24] 等。

为了更好地说明 PCA 数据压缩方法在光谱数据挖掘中的应用原理和过程，下面较详细地介绍李乡儒等人 [19] 的研究工作，其属于有监督数据挖掘。他们探讨了星系与类星体光谱分类，训练集合中的每条光谱均有类别标示，其目标是按照既定的标准对其他的观测数据快速准确地进行分类归档，考虑的波长范围是 3 800 ∼ 9 000 Å 。由于每条光谱有 3 791 个采样点，数据维数较高，观测光谱往往受到多种噪声干扰，而且不同流量之间往往有一定的相

关性，这些因素会导致分类工作的计算效率较低和过学习。因此，他们首先对训练数据进行 PCA 分析 (公式 (1) 和公式 (2)), 得到主成分方向 $\{v_1, \cdots, v_{3791}\}$ 。如果假定训练样本和待处理的观测数据独立同分布，则通过式 (5) 的 PCA 特征变换后各数据成分满足：

$$cov(x^T v_i, x^T v_j) = v_i^T E[(x - E(x))(x - E(x))^T] v_j = v_i^T \Sigma_P v_j \approx v_i^T \hat{\Sigma}_P v_j = \delta(i-j)\lambda_i, \quad (6)$$

即数据成分之间的相关性被消除了，其中，$\delta(x)$ 是狄利克雷函数，当 $x = 0$ 时函数值为 1，否则函数值为 0。而且，按照主成分分析的基本原则和模型假设，噪声等干扰因素的影响往往较小，并集中于方差 λ_i 较小的方向 v_i。因此，通过在 PCA 特征变换式 (5) 中取 $k = 4 < 3\,791$ 能够一定程度上减小噪声的干扰。然后，在投影后的四维 PCA 特征空间中对光谱进行分类。

关于主成分分析在特征谱构造中的应用将在本文第 3 章介绍。需要注意的是，PCA 是一种线性方法, 为了处理光谱数据本身的非线性特点，文献 [11, 25, 26] 采用了对光谱数据进行分区分析的方法，其思想是曲线拟合理论中的局部线性化。另外，在光谱模式分析中，有静态 [13, 25, 26] 和动态 [19] 两种使用 PCA 的方案：“静态” 是指在运用 PCA 方法之前首先将光谱移至静止波长，剔除红移因素的影响；在 “动态” 方案中，则不剔除红移的影响，PCA 的作用主要是数据压缩、提高计算效率和抑制噪声的负面影响。在特征谱构造中需要使用静态 PCA；在基于 PCA 的数据压缩中，可根据研究目标选择合适的实现方案。

PCA 是一种高效的数据降维方法，易于使用，且去除了因子之间的相关性。其局限性是，这是一种全局分析工具，在时间/波长轴上没有分辨能力和定位功能，这一方面会导致基于该方法的光谱数据挖掘效果在有些情况下会较差一些，另一方面，通过它不能对 “时” 进行分辨 (在光谱分析中,“时” 是指波长)，由此导致其特征的物理意义一般难于分析；而且，其分析结果容易受个别离群数据影响。

2.2 小波变换

小波变换 (Wavelet Transform) 是一种有效的时频分析工具，在光谱特征提取中亦得到了较为广泛的关注和研究。如果将信号看作是时间的函数，小波 $\psi(t)$ 是平方可积，均值为零的函数：

$$\int_{-\infty}^{+\infty} \psi^2(t)\mathrm{d}t = 1, \quad (7)$$

$$\int_{-\infty}^{+\infty} \psi(t)\mathrm{d}t = 0, \quad (8)$$

且能量集中在以 $t = 0$ 为中心的邻域内。对小波 $\psi(t)$ 做伸缩和平移后可得到一族时频原子

$$\psi_{u,s}(t) = \frac{1}{\sqrt{s}}\psi(\frac{t-u}{s}). \quad (9)$$

任一信号 $f(t)$ 关于时间 u、尺度 s 的小波变换为

$$Wf(u,s) = \int_{-\infty}^{+\infty} f(t)\psi_{u,s}^*(t)\mathrm{d}t = \frac{1}{2\pi}\int_{-\infty}^{+\infty} \hat{f}(w)\hat{\psi}_{u,s}^*(w)\mathrm{d}w, \quad (10)$$

其中，$\hat{f}(w)$ 和 $\hat{\psi}^*_{u,s}(w)$ 分别是 $f(t)$ 和 $\psi^*_{u,s}(t)$ 的傅里叶变换，$\psi^*_{u,s}(t)$ 是 $\psi_{u,s}(t)$ 的复共轭。如果小波 $\psi(t)$ 是解析的，且其傅里叶变换 $\hat{\psi}(w)$ 仅在以某个 $\eta > 0$ 为中心的局部区域内非零，则 $\psi_{u,s}(t)$ 和 $\hat{\psi}^*_{u,s}(w)$ 的能量便分别集中在时间轴和频率轴上中心为 u 和 η/s，宽度为与 s 和 $1/s$ 成正比的某个局部区域上。从而，根据式 (9)，小波特征 $Wf(u,s)$ 仅仅依赖于信号 $f(t)$ 及其傅里叶变换 $\hat{f}(w)$ 在 $\psi_{u,s}(t)$ 和 $\hat{\psi}^*_{u,s}(w)$ 能量集中的某个局部时频区域上。所以，小波能够在时间和频率两方面同时实现局部化，检测和提取光谱中的不同波长位置辐射能量的瞬态变化。在天体光谱的分析中，时间对应于波长，小波的时间、频率局部化有助于我们分析不同波段、不同频率的天体辐射对当前分析的重要性，有利于对分析结果物理内涵的探索。上述介绍主要基于文献 [27, 28]，关于小波基本原理的更详尽阐述和相关软件包的介绍请参考文献 [27–29]。

主成分分析是以整个观测光谱为分析单元，而小波分析则是以光谱中的局部流量分量为分析单元，适合于分析信号中的噪声情况、连续谱估计和局部结构检测。例如，Starck 等人[30]探讨了噪声和连续谱的估计；Madgwick 等人 [31] 基于àtrous 小波研究了噪声的抑制及其在超新星搜索中的应用，Fligge 和 Solanki[32] 基于小波包研究了光谱噪声的剔除。邢飞和郭平通过 Mallat 小波算法对光谱信号进行分解，并将分解后的高频子空间作为噪声成分丢弃，实现光谱信号的特征提取，然后使用支持向量机方法研究了恒星光谱的分类 [33]；刘蓉和段福庆等人基于小波变换技术研究了星系光谱的分类 [34]，该工作首先使用 Spline2 小波对光谱信号进行分解，并以第四层小波系数作为光谱谱线信息的描述，然后在谱线描述空间中使用主成分分析方法对数据进行约简，最后利用 Fisher 线性判别分析方法实现正常星系和活动星系光谱的分类；张怀福、赵瑞珍和罗阿理 [35] 通过使用小波包提取光谱特征研究了活动天体光谱和非活动天体光谱的分类，他们首先使用 Daubechies 3 小波对天体光谱进行三层小波包分解，并以光谱在第三层中各个频率子空间中的能量累积和作为光谱分类特征的描述，形成一个 8 维小波包特征空间，然后在该空间中使用支持向量机方法对红移未知的活动天体光谱和非活动天体光谱进行分类。刘中田、赵瑞珍、赵永恒和吴福朝 [36] 基于 Daubechies 小波研究了晚型星光谱的识别：在该工作中，首先使用 Daubechies 小波对光谱信号进行频谱分解，并以第五层小波系数作为分类描述特征，然后通过分析光谱描述特征的能量峰值情况给出了晚型星的识别方法。

如上所述，基于小波变换的特征提取步骤一般是：首先对光谱进行小波分解；然后，将其中的高频成分作为噪声丢弃，并通过交互方式选择某 (几) 个频段的小波系数作为特征，据此进行光谱的分类、参数估计等数据挖掘工作。小波分析方法实现了特征信息的局部化，但是其难点是尺度和位置的选择 [37]。上述研究的局限性是：(1) 尺度的选择未能实现自动化，均是通过人工方式进行选择；(2) 尺度的选择未能实现随时间/波长位置自适应变化，但是，不同类型的光谱和同一类型光谱的不同波长位置的特征信息往往存在尺度差异；(3) 虽然小波方法能够在表征上实现对位置的区分，但上述研究并未探讨其选择问题，即放弃了小波的“时” 分析能力，因此在使用中本质上亦成为一种全局方法。

2.3 流形学习

流形学习 (manifold learning) 是一种有效的数据降维和本质结构抽取方法。基本思想

是，观测数据一般仅受少数几个因素影响，而且常常是平滑变化，甚至可以局部线性近似，因此，可以假设它们近似落在一个流形上，从而可以通过流形学习的方法将这些高维数据映射到对应的内在本质低维结构上。最近几年，以 LLE(Locally Linear Embedding)[38]，ISOMAP(Isometric Mapping) [39] 和 Laplacian Eigenmaps[40] 为代表的流形学习方法得到了广泛的关注和研究, 它们均是通过保持流形的某些局部属性实现降维。例如，LLE 的假设是每个数据点均可用它的近邻点线性表示，在低维空间和高维空间中的权值不变；MDS(Multidimensional scaling) 的基本思想是数据维数约简前后任意两点间的距离不变[41]；ISOMAP 则是建立在 MDS 基础之上，力求保持两两数据点间的测地距离；Laplacian 的基本假设是，在高维空间中离得很近的点投影到低维空间中后也应该离得很近。

目前，该类方法在光谱数据分析中的应用已经得到了一定的关注和研究。例如，Vanderp–las 和 Connolly[42] 研究了 LLE 特征提取方法在光谱分类中的应用，结果表明，他对于发射线天体光谱的分类效果较好; 许馨、吴福朝、胡占义和罗阿理基于 LLE 特征提取方法研究了正常星系光谱的红移估计问题 [43]; Richards 等人探讨了基于 Diffusion Maps 方法的天体光谱特征提取，并据此研究了特殊天体光谱的识别和星系光谱的红移估计 [44]，该方法保持的流形属性是数据之间的连通性 (connectivity); Daniel 和 Connolly 等人研究了基于 LLE 数据约简的恒星光谱分类 [45]，研究结果表明，大部分恒星光谱能够用三维空间中的一个序列表征，而且，偏离该序列的光谱一般是强发射线天体 (如错分类的星系) 和宽吸收线天体 (如碳星)。

与主成分分析方法一样，流形学习方法也是一种全局分析工具，以整个观测光谱为分析单元，其局限性是既不能对 “时” 进行分辨，亦不能对 “频” 进行分析，由此导致该方法对光谱的流量定标畸变和波长定标畸变较为敏感。

2.4 有监督特征提取

与上述特征提取方法不同，有一类方法在其目标函数中融合了后续处理的具体需要，这类方法称为有监督特征提取方法，相应地，前述方法称为无监督方法。Bailer-Jones 和 Fiorentin 等人通过神经网络有监督方法研究了恒星有效温度、表面重力、化学丰度的估计[10, 46]。文献 [9] 通过变宽核回归方法研究了恒星大气参数的估计。许馨等人通过将 Fisher 判别分析方法与核技巧结合起来研究了恒星、星系和类星体光谱的分类问题，并称该方法为广义判别分析方法 (Generalized Discriminant Analysis, GDA)[47]。在该方案中，首先通过核方法诱导出的非线性映射将光谱数据映射到某个高维空间 F 中, 然后在空间 F 中进行线性判别分析。杨金福等人提出了一种基于核技巧的覆盖算法——核覆盖算法，并研究了他在正常星系、恒星、星暴星系和活动星系核光谱分类中的应用 [48]。该算法将核技巧与覆盖算法相结合, 并在特征空间中抽取支持向量, 这里的支持向量即是提取的分类特征。李乡儒等人[49, 50]结合类星体光谱、星系光谱、Seyfert 1 光谱和 Seyfert 2 光谱的分类问题研究了基于线性判别分析 (Linear Discrimitive Analysis, LDA) 和相关向量机 (Relevance Vector Machine, RVM) 的光谱特征提取，并探讨了光谱识别中有监督特征提取的必要性和重要性。

与无监督特征提取方法相比，有监督特征提取方法所获得的特征一般更有针对性，整体效果较好，但其难点是符合需求的目标函数的设计。另外，这两类方法的适用情况也有一些差异，无监督方法适于探索分析，重在求新，发现新的规律；有监督方法适合于按照既定的

规范自动进行归档分类或参数估计，重在求精，即准确性。

2.5 统计约简法的综合比较

为了清晰明了，我们对上述 4 类统计约简法做了综合比较 (如表 1 所示)。

表 1 光谱特征提取的统计约简方法的特点分析与比较

方法	时频分辨能力	特征选择能力	分析粒度	对离数据的稳健性	适用场景	物理意义的可解释性
主成分分析	无	有	光谱空间	差	探索式挖掘	中
小波变换	有	无	流量空间	好	探索式挖掘	优
流形学习	无	无	光谱空间	差	探索式挖掘	差
有监督方法	一般无	有	光谱空间	与具体方法有关	归档式分类、参数估计	差

3 特征谱法

特征谱是指通过某种方法构造的人工“光谱”，可用于近似表征某些类型的天体观测光谱。有两类特征谱构造方法，其一是强调某些光谱频谱特征的准确表征，例如，发射线和幂律谱特点等。这方面的工作如，Vanden Berk 和 Richards 等人基于观测光谱流量的中值法和几何均值法研究了类星体特征谱的构造 [51]。第二类方法则强调对观测光谱的近似表达能力。例如，在基于 PCA 的特征谱构造中，假设 $v_1,\cdots,v_m$ 是其中的特征向量 (式 (2)), 则它们形成了光谱空间的一组正交基，于是，对于任意一条光谱 x，可表征为特征向量的线性组合：

$$x=\sum_{i=1}^{m}(x^Tv_i)v_i, \tag{11}$$

而对于选定的某个 $1\leqslant k\leqslant m$(式 (4))，可用下式对光谱做近似：

$$x\approx\sum_{i=1}^{k}(x^Tv_i)v_i. \tag{12}$$

相关工作如，McGurk 等人根据 Sloan 发布的 98 063 条观测光谱，运用 PCA 方法研究了恒星特征谱的构造，该结果可应用于新观测光谱的识别和光谱分类器的训练 [25]；Yip 等人针对 Sloan 观测的 16 707 条类星体光谱研究了 PCA 特征谱的构造，并据此探讨了类星体光谱的分类，结果表明该分类问题与天体的红移和光度相关 [26]; Eisenstein 和 Hogg 等人研究了基于主成分的星系平均谱构造 [52]; Madgwich 等人研究了星系特征谱的构造及其在超新星搜索中的应用 [31]。Connolly 等人研究了基于主成分分析的星系特征谱的构造，以及二维特征谱空间中的星系光谱分类，星系的光谱分类与形态分类的关系 [53]。基于 PCA 的特征谱构造实际上是以方差贡献率较高的数个特征向量张成的子空间作为某类光谱所在空间的近似，从而，每条相关的观测光谱均可用选定的 PCA 特征谱的线性组合近似。

以特征谱作为某类天体光谱的模版，可用于观测光谱的识别、红移估计、光谱成分分解等。例如，McGurk 等人基于分段 PCA 研究了恒星光谱的主成分特征谱的构造及其在合成光谱构造中的应用 [11]；Li 等人研究了恒星星系 PCA 特征谱的构造，并据此探讨了星系光谱中恒星成分和发射线成分的分解 [54]; Vanden Berk 等人基于星系和类星体 PCA 特征谱研究了宽线活动星系核光谱中寄主星系成分的分离，并探讨了分解精度的影响因素 [55]; 屠良平等人基于特征谱研究了超新星光谱的识别 [56]，该工作首先在超新星候选范围缩减中利用了 PCA 特征谱，然后在超新星光谱的进一步筛选中，使用实测光谱作为特征谱；Duan 等人基于特征谱的匹配研究了天体光谱的自动分类 [57]，其中使用的恒星特征谱和正常星系特征谱均由 PCA 方法构造，类星体特征谱选用的是 Vanden Berk 等人构造的光谱模版 [51]；许馨、罗阿理、吴福朝和赵永恒以 Kinney 等人构造的光谱模版作为特征谱 [58]，研究了正常星系光谱的红移估计问题 [59]；Bai 和 Guo 以 EKF 平滑后的光谱作为特征谱，据此使用径向基神经网络研究了恒星光谱的分类 [60]。Johnston 和 Richards 等人基于星系和类星体的特征谱研究了光谱中星系和类星体成分的分解 [61]。

基于特征谱的方法存在 3 方面的问题：(1) 往往计算量较大，效率较低；(2) 高质量特征谱的构造，特别是存在红移、天光污染和其他噪声影响的情况下，往往需要大量的已分类数据，这在有些情况下难于满足；(3) 如果某些类型光谱的次型较多，且相应的观测数据较少，会导致所构造的特征谱的代表性较差，进而导致据此所做的模式分析结果有较大偏差或遗漏。

4 谱线法

因为谱线是天体光谱中最显著的特征，所以基于它的光谱模式分析得到了广泛的关注和研究。例如，Dessauges-Zavadsky 等人 [62] 基于低电离发射线

$$([\mathrm{OI}]\lambda 6300, [\mathrm{NII}]\lambda 6584, [\mathrm{SII}]\lambda\lambda 6717, 6731),$$

氧线

$$([\mathrm{OI}]\lambda 6300, [\mathrm{OII}]\lambda 3727, [\mathrm{OIII}]\lambda 5007),$$

以及线强比

$$R_{23} = ([\mathrm{OII}]\lambda 3727 + [\mathrm{OIII}]\lambda 4959 + [\mathrm{OIII}]\lambda 5007)/\mathrm{H}\beta,$$

和诊断图法研究了发射线星系光谱的分类，并给出了一种效果较好的诊断图 $\lg R_{23}$ vs. $\lg[\mathrm{OI}]\lambda 6300/\mathrm{H}\alpha$. Kewley 等人基于 Sloan 发布的 85 224 条发射线星系光谱，以及 $[\mathrm{OIII}]/\mathrm{H}\beta$, $[\mathrm{NII}]/\mathrm{H}\alpha$, $[\mathrm{SII}]/\mathrm{H}\alpha$, $[\mathrm{OI}]/\mathrm{H}\alpha$ 和 $\lg[\mathrm{OIII}]/\mathrm{H}\beta$ 等谱线比研究了星系光谱的分类 [63]。当红移高于 0.4 时，$[\mathrm{NII}]\lambda 6717 + 6731$ 和 $\mathrm{H}\alpha$ 等发射线将逐渐超出观测波长范围，这会导致基于这些发射线的线比分类法失去效用。为此，Lamareille 等人研究了基于谱线的高红移发射线星系光谱的分类问题 [64]。

在基于谱线的海量光谱模式分析中，谱线的自动提取和描述是其关键问题。相关工作如，在 Sloan[65, 66] 和文献 [30] 中采用àtrous 小波提取光谱谱线，文献 [67] 和 [68] 研究了 Lyα 线丛中吸收线线宽等相关属性。由于郭守敬望远镜 (LAMOST) 项目数据处理需求的推动，近几年国内对特征谱线的自动提取、描述和应用做了广泛的研究。例如，罗阿理和赵永恒使用小波方法研究了恒星、近邻星系和 AGN 天体光谱的谱线自动提取 [69]，该工作首先使用小波方法将光谱中的连续谱去除，然后在小波域使用隐马尔可夫方法估计光谱中的噪声分布并进行降噪处理，最后据此采用局部阈值方法和高斯拟合技术提取谱线并测量其线心波长值等参数。赵瑞珍等人提出了一种基于稀疏表示的谱线自动提取方案 [70]：他们首先用基于稀疏表示的小波方法去除噪声；然后利用小波变换与样条拟合相结合的方法拟合出光谱的伪连续谱，并据此对光谱进行归一化处理；最后，通过对归一化后的光谱设置自适应局部阈值来提取谱线。段福庆和刘蓉等人研究了均值漂移算法在光谱滤波和谱线提取等方面的应用 [71, 72]，研究结果表明该方法能够去除脉冲噪声，抑制非脉冲噪声、天光背景噪声和随机噪声，具有较强的谱线保护能力，整体上优于小波硬阈值法、高斯滤波和中值滤波方法。张继福等人研究了通过对谱线波峰强度、峰宽和形状信息离散化的方法描述光谱 [73]，继而探讨了基于该描述的天体光谱离群数据的发现，其中波峰强度分为强、一般、弱和无 4 种情况，峰宽分为窄和宽两种情况，谱线形状分为吸收线和发射线两种情况。Qiu 等人通过对光谱形态细节的研究，提出了 6 种描述谱线的基元：峰基元 (p)、峰左基元 (pl)、峰右基元 (pr)、谷基元 (v)、谷左基元 (vl) 和谷右基元 (vr)，并据此给出了一种对光谱整体做出完整描述的语言——光谱描述语言，同时探讨了他在恒星光谱分类中的应用 [74]。Duan 等人 [57] 基于谱线的匹配研究了正常星系和类星体光谱红移的估计，并继而探讨了基于特征谱匹配的光谱识别。

谱线法的突出优点是物理意义强，容易解释，局限性是：(1) 谱线是一个高级的认知概念，它的可靠提取和正确认证需要复杂的人类知识指导，所以，基于对光谱数据低层处理的谱线自动提取方法及其应用的可靠性受光谱频谱信息的复杂度和质量影响较大 [74]；(2) 谱线的描述受所用仪器、波长和流量标定情况影响较大，例如，谱线形状、谱线强度、半高全宽等；(3) 基于谱线的方法，特别是线比法，受红移范围限制明显，在低红移光谱中出现的谱线，在高红移目标中会移出观测波长范围 [64]。

5 结 论

如前所述，特征提取是对天体辐射能量测量指标的分解、重组和选择的过程，关键环节有：(1) 特征的检测和定位；(2) 特征的表达。按照特征的表达方式，已有的光谱特征提取方法本质上可分为统计约简法、特征谱法和谱线法。在特征提取方法的选择上，需要考虑的问题有: 挖掘结果物理意义的可解释性，自动处理的效率，对噪声和畸变的稳健性，以及适用性等。例如，统计约简法的优点是一般均有比较自动化的步骤，易于操作和使用，但是如果使用的时候不考虑光谱处理的科学问题需求，则易于陷入从输入到输出的纯 “黑盒” 式数字游戏，导致结果失去物理意义；谱线法的典型优点是物理意义强，基于谱线法的光谱挖掘研究

易于集成天文学家的先验知识，缺点是谱线提取的稳健性易受噪声和畸变影响，线比分类法的适用性往往受光谱红移范围限制。

特征的检测和定位是指特征提取方法的时、频分辨能力和自动选择能力，它关系到特征提取方法的数据压缩效率和物理意义的可解释性。例如，谱线法具有较好的时分辨和选择能力，所以，数据压缩效果和基于此的自动挖掘方法的效率也较高。另外，我们在本文第 2 章至第 4 章中亦从特征的检测和定位角度对文献中的已有方法作了分析，并将各类方法的优缺点简要总结至表 2。

表 2　文献中已有光谱特征提取方法的特点

光谱表达方式	表达模式	优点	局限性
统计约简法	全局	易于使用	1) 物理意义一般比较难于解析；2) 在时间/波长轴上没有自动定位和选择的功能；3) 受观测仪器、红移以及流量标定和波长标定情况影响较大。
特征谱法	全局	物理意义强	1) 代表性差，特别是, 如果某类天体的观测光谱较少，子型和次型多样, 红移变化范围大，且有些互相差异很大，则难于构建高质量和高代表性的特征谱，从而影响模式分析的准确性 [56]；2) 特征谱往往是上千维的，由此导致基于特征谱的模式分析方法效率较低; 3) 受观测仪器、红移以及流量标定和波长标定情况影响较大。
谱线法	半全局	物理意义强	1) 自动提取和认证的可靠性差；2) 描述的准确性受流量标定和波长标定情况影响，且谱线形状亦随红移而变化；3) 相应的模式分析方法，特别是线比法，受红移变化范围影响较大。

参考文献：

[1] Szalay A, Gray J. Science, 2001, 293: 2037
[2] 张彦霞，赵永恒，崔辰州. 天文学进展, 2002，20(4): 312
[3] 李丽丽，张彦霞，赵永恒，杨大卫. 天文学进展, 2006, 24(4): 285
[4] Ball N M，Brunder R J. International Journal of Modern Physics D (IJMPD), 2010, 19(7): 1049
[5] Liu H, Hiroshi M, Rudy S, Zhao Z. Journal of Machine Learning Research, 2010, W&P 10: 4
[6] 李乡儒, 胡占义, 赵永恒, 刘中田. 天文学报, 2007, 48(3): 280
[7] Li X R, Hu Z Y, Zhao Y H, Liu Z T. Chinese Astronomy and Astrophysics, 2008, 32(1): 13
[8] Jolliffe I T. Principal Component Analysis, second edition. New York: Springer-Verlag, 2002
[9] Zhang J N, Wu F C, Luo A L, Zhao Y H. Chinese Astronomy and Astrophysics, 2006, 30(2): 176
[10] Fiorentin P R, Bailer-Jones C A L, Lee Y S, et al. A&A, 2007, 467: 1373
[11] McGurk R C, Kimball A E, IvezićŽ. AJ, 2010, 139(3): 1261
[12] Singh H P, Gulati R K, Gupta R. MNRAS, 1998, 295: 312
[13] Yip C W, Connolly A J, Szalay A S, et al. AJ, 2004, 128(2): 585
[14] Folkes S, Ronen S, et al. MNRAS, 1999, 308(2): 459

[15] Galaz G, Lapparent V. A&A, 1998, 332:459
[16] Ronen S, Aragon-Salamanca A, Lahav O. MNRAS, 1999, 303: 284
[17] Madgwick D S, Coil A L, et al. ApJ, 2003, 599: 997
[18] Rogers B, Ferreras I, et al. MNRAS, 2007, 382: 750
[19] 李乡儒，卢瑜，周建明，王永俊. 光谱学与光谱分析, 2011, 31(9): 2582
[20] Pâris I, Petitjean P, Rollinde E, et al. arXiv1104.2024, 2011
[21] Suzuki N, Tytler D, Kirkman D, et al. ApJ, 2005, 618: 592
[22] Lee K G, Suzuki N, Spergel D N. arXiv:1108.6080v1, 2011
[23] Suzuki N. ApJS, 2006, 163(1): 110
[24] Boroson T A, Green R F. ApJS, 1992, 80: 109
[25] McGurk R C, Kimball A E, et al. AJ, 2010, 139(3): 1261
[26] Yip C W, Connolly A J, Vanden Berk D E, et al. AJ, 2004, 128(6): 2603
[27] Mallat S. A Wavelet Tour of Signal Processing, 2nd ed. San Diego: Academic Press, 1999
[28] Mallat S. A Wavelet Tour of Signal Processing, 3rd ed. San Diego: Academic Press, 2008
[29] Daubechies I. Ten lectures on wavelets. Philadelphia:Society for Industrial and Applied Mathematics (SIAM), 1992
[30] Starck J L, Siebenmorgen R, Gredel R. ApJ, 1997, 482: 1011
[31] Madgwick D S, Hewett P C, et al. ApJ, 2003, 599: L33
[32] Fligge M, Solanki S K. Astron. Astrophys. Suppl. Ser., 1997, 124(3): 579
[33] 邢飞, 郭平. 光谱学与光谱分析, 2006, 26(7): 1368
[34] 刘蓉, 段福庆，刘三阳，吴福朝. 电子学报, 2005，33(11): 2059
[35] 张怀福, 赵瑞珍, 罗阿理. 北京交通大学学报, 2008, 32(2): 30
[36] 刘中田, 赵瑞珍, 赵永恒, 吴福朝. 光谱学与光谱分析, 2005, 25(7): 1158
[37] Huang K, Aviyente S. IEEE Trans. Image Proc., 2008, 17(9): 1709
[38] Roweis S T, Saul L K. Science, 2000, 290: 2323
[39] Tenenbaum J B, de Silva V, Langford J C. Science, 2000, 290: 2319
[40] Belkin M, Niyogi P. Neural Computation, 2003, 15(6): 1373
[41] Borg I, Groenen P. Modern Multidimensional Scaling: theory and applications, 2nd ed. New York: Springer-Verlag, 2005
[42] Vanderplas J, Connolly A. AJ, 2009, 138(5): 1365
[43] 许馨, 吴福朝, 胡占义, 罗阿理. 光谱学与光谱分析, 2006, 26(1): 182
[44] Richards J W, Freeman P E, Lee A B, Schafer C M. ApJ, 2009, 691: 32
[45] Daniel S F, Connolly A J, Schneider J, et al. arXiv:1110.4646, 2011
[46] Bailer-Jones C A L. A&A, 2000, 357: 197
[47] 许馨，杨金福, 吴福朝，赵永恒. 光谱学与光谱分析，2006, 26(10): 1960
[48] 杨金福，许馨，吴福朝，赵永恒. 光谱学与光谱分析，2007, 27(3): 602
[49] 李乡儒, 胡占义, 赵永恒, 李晓明. 光谱学与光谱分析，2009, 29(06): 1702
[50] 李乡儒, 胡占义, 赵永恒. 光谱学与光谱分析, 2007, 27(9): 1898
[51] Vanden Berk D E, Richards G T, et al. AJ, 2001, 122: 549
[52] Eisenstein D J, Hogg D W, et al. ApJ, 2003, 585: 694
[53] Connolly A J, Szalay A S, Bershady M A, et al. AJ, 1995, 110(3): 1071
[54] Li C, Wang T G, Zhou H Y, et al. AJ, 2005, 129: 669
[55] Vanden Berk D E, Shen J, Yip C W, et al. AJ, 2006,131: 84
[56] 屠良平，罗阿理，吴福朝，赵永恒. 中国科学: 物理学 力学 天文学, 2010，40(10): 1282
[57] Duan F Q, Liu R, Guo P, et al. Research in Astron. Astrophys, 2009, 9(3): 341
[58] Kinney A L, Calzetti D, Bohlin R C, et al. ApJ, 1996, 467: 38
[59] 许馨，罗阿理，吴福朝，赵永恒. 光谱学与光谱分析, 2005, 25(6): 996
[60] Bai L, Guo P, Hu Z Y. The Chinese Journal of Astronomy and Astrophysics, 2005, 5(2): 203
[61] Johnston D E, Richards G T, et al. ApJ, 2003, 126: 2281

[62] Dessauges-Zavadsky M, et al. A&A, 2000, 355: 89
[63] Kewley L J, Groves B, Kauffmann G, et al. MNRAS, 2006, 372(3): 961
[64] Lamareille F. arXiv:0910.4814, 2011
[65] Stoughton C, Lupton RH, et al. AJ, 2002, 123: 485
[66] http : //www.sdss.org/dr5/algorithms/redshift_type.html, 2011
[67] Theuns T, Zaroubi S. MNRAS, 2000, 317: 989
[68] Meiksin A. MNRAS, 2000, 314: 566
[69] 罗阿理, 赵永恒. 天体物理学报, 2000, 20(4): 427
[70] 赵瑞珍, 王飞, 罗阿理, 张彦霞. 光谱学与光谱分析, 2009，29(7): 2010
[71] 段福庆, 周明全, 张家才. 吉林大学学报 (工学版), 2007, 37(3): 634
[72] 刘蓉, 段福庆, 刘三阳, 吴福朝. 电子与信息学报, 2006, 28(2): 312
[73] 张继福, 蒋义勇, 胡立华, 蔡江辉, 张素兰. 自动化学报, 2008，34(9): 1060
[74] Qiu B, Hu Z Y, Zhao Y H. SPIE, Seattle, July 7-11, 2002

Feature Extracting Methods in Spectrum Data Mining

LI Xiang-ru

(School of Mathematical Sciences, South China Normal University, Guangzhou 510631, China)

Abstract: Feature extraction is the fundamental step in spectrum data mining, which determines both the quality of the mining results and the efficiency, robustness, complexity of the mining system. This work reviews the current state of celestial spectrum feature extracting methods, introducs the fundamental ideas, analyzes their superiorities, limitations and applicabilities. By extracting features, the measurements of a spectrum are decomposed, reorganized and selected. Based on the characteristics of information expression, we classify the available feature extraction methods into three categories: statistical reduction method, characteristic spectrum method, and spectral line method. Their applications in spectrum data mining are also introduced. For clarity, the statistical reduction method is further classified into the following four classes: principal component analysis (PCA), wavelet transform (WT), manifold learning and supervised methods. In addition, we also study such characteristics of these methods as time-frequency analysis, the interpretability of physical meaning, robustness to calibration distortion, robustness to outlier, etc.

Key words: spectrum; data mining; feature extraction