第27卷，第3期
2007年3月

光 谱 学 与 光 谱 分 析
Spectroscopy and Spectral Analysis

Vol. 27, No. 3, pp606-609
March, 2007

面向 LAMOST 的天体光谱离群数据挖掘系统研究

张继福[1,2]，蔡江辉[1]

1. 太原科技大学计算机科学与技术学院，山西 太原 030024
2. 中国科学院自动化研究所模式识别国家重点实验室，北京 100080

摘 要 在宇宙中寻求未知天体是人类探索宇宙奥妙所追求的目标之一，离群数据挖掘是发现未知天体光谱数据的一种有效途径。文章首先以 VC++和 Oracle9i 为开发工具，设计与实现了面向 LAMOST 的恒星光谱离群数据挖掘系统，并给出了其软件体系结构和模块功能。其次，对基于中值滤波器的恒星光谱数据预处理、基于距离的恒星光谱数据聚类、基于距离支持度的恒星光谱数据离群数据挖掘、基于主分量分析法 PCA 的恒星光谱数据离群数据的三维可视化等主要关键技术进行了详细描述。最后，基于 SDSS 恒星光谱数据的运行结果表明，利用该系统寻找天体光谱离群数据是可行的，从而为寻找未知的、特殊的天体光谱数据提供了一种新途径。

关键词 天体光谱数据；离群数据；聚类；距离支持度

中图分类号：TP311 **文献标识码**：A **文章编号**：1000-0593(2007)03-0606-04

引 言

目前我国正在建造一台大天区面积多目标光纤光谱望远镜(简称 LAMOST)，是国家重大科学工程项目，总投资达 2.35 亿人民币，也是世界上光谱获取率最高的望远镜。由于 LAMOST 具有以较高效率大规模测量天体光谱的能力，可提供的研究课题将从银河系、星系、星系团、活动星系核，直到宇宙大尺度结构。预计从 2006 年底起，每个观测夜将收集 2 到 4 万条光谱的数据，LAMOST 所观测到的光谱数据容量将达 4TB[1]。如此庞大的数据，利用传统人工数据分析方式将无法满足实际需求，因此，急需一种以计算机为主的数据分析技术来解决这一问题。

当前天体光谱数据分析与处理主要集中在光谱的分类和识别方面，天文学界研究较多的是恒星光谱的分类识别，具有代表性的是 Autoclass，它是基于贝叶斯统计的一种分类方法，其独特的分类结果发现了一些以前未注意到的光谱类型和谱线。Gulati 等首先采用两层 BP 神经网络方法，用于恒星光谱次型的分类。Jones 等采用多个 BP 网络进行恒星光谱次型的分类识别。薛剑桥等采用自适应神经网络 SOFM 进行了恒星光谱的分类识别。邱波等采用基于粗糙集的自动提取分类规则的方法进行了恒星光谱的分类识别。覃冬梅等提出了基于主分量分析法的二维恒星特征空间的快速光谱识别方法等[1-4]。由于天文界对宇宙的认识还比较有限，LAMOST 巡天计划的一个重要任务是要发现一些新的、特殊类型的天体，因此，如何利用数据挖掘技术从海量天体光谱数据中发现未知的、特殊的天体及天体规律是数据挖掘值得研究和探索的新应用领域。

离群数据(孤立点)挖掘[5]，是数据挖掘研究的一个重要分支。离群数据是指明显偏离其他数据、不满足数据的一般模式或行为、与其他数据不一致的数据。由于人类对宇宙的认识还比较有限，因此利用离群数据挖掘技术在海量天体光谱数据中寻找稀有的未知类型天体光谱数据，对于人类探索宇宙奥妙具有重要的实际意义。本文研究的面向 LAMOST 的天体光谱离群数据分析系统，已于 2005 年 11 月份通过了国家科技部的验收，并在中科院国家天文台采用 SDSS 恒星数据进行了试运行，取得了较好的效果，为寻找未知的、特殊的天体光谱数据提供了一条新途径。

1 系统的软件体系结构及功能

天体光谱离群数据挖掘系统的主要目的是在海量天体光谱数据中，通过大范围的、无偏差的多波段的探测，找到奇异天体光谱数据，以期发现某些特殊的、未知的天体。天体

收稿日期：2005-12-05，**修订日期**：2006-04-21

基金项目：国家自然科学基金项目(60573075)，国家“863”高技术研究发展计划基金项目(2003AA133060)和山西省自然科学基金项目(2006011041)资助

作者简介：张继福，1963 年生，太原科技大学教授，中国科学院自动化研究所博士后 e-mail：jifuzh@sina.com

光谱数据是海量高维数据，数值变化范围非常大，直接运算效率很低，同时某些非数值型数据无法直接参加运算，所以系统首先要对天体光谱数据进行预处理。离群数据挖掘采用的是基于聚类的方法，如何提高聚类的运算速度直接影响整个系统的运行效率。为了能够高效、准确地处理海量高维数据，系统中采用一种新的聚类算法 DB-HDLO，从而可以在聚类的基础上进行离群数据挖掘。该方法能够根据天文学家的不同要求发现离群数据，系统通过对一定约束条件的修改实现了这一功能。数据挖掘的结果表示是另一个值得关心的问题，为了实现这一功能，通过主分量分析法(PCA)[6]进行降维，将高维天体光谱数据映射到三维空间，并将结果可视化输出。基于上述讨论，该系统的功能模块如图 1 所示。

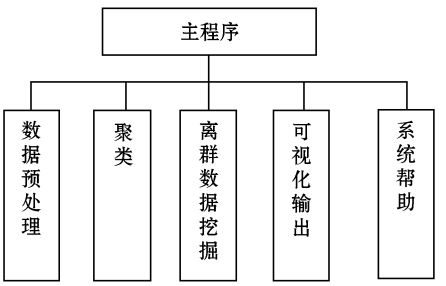

Fig. 1　Function modules

图 2 是该挖掘系统的软件体系结构。先通过用户接口输入数据归一化的参数设置值，由预处理技术将天体光谱数据归一化。归一化之后的天体光谱数据，通过聚类、离群数据挖掘模块的处理，将发现离群天体光谱数据，最后将结果降维，生成可视化最终结果，并由用户接口进行输出。

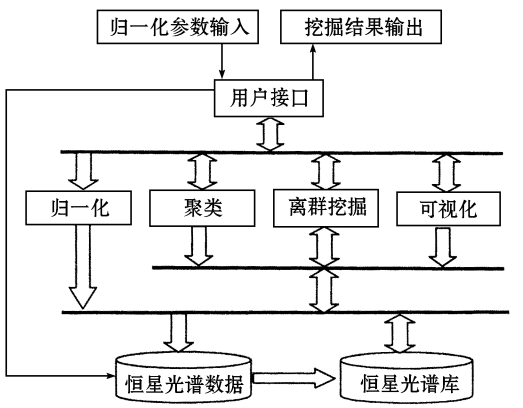

Fig. 2　Software architectural structure

基于上述功能和其软件体系结构，采用 VC＋＋6.0 和 Oracle9i 为开发工具，设计与实现了天体光谱离群数据挖掘系统。

2　关键技术

2.1　光谱数据预处理

在 LAMOST 收集到的数据中，一条天体光谱由连续谱、特征谱线以及噪声组成，而连续谱形状对恒星来说是重要的分类特征，在恒星分类中起关键作用。同时，天体光谱的原始数据是由每一个波长对应的流量和光谱的物理化学性质组成，流量值变化范围很大，可达 $10^{-19}\sim10^{6}$，这将大大影响运算效率，所以在该系统中预处理的主要工作是提取连续谱，缩小数据量值。

关于连续谱的提取，用多项式逼近是常用的方法之一，同时还有许多别的方法，如形态滤波器、小波变换、中值滤波器等。形态滤波器是一种对信号的几何特征进行变换的非线性滤波器，但实现较复杂，而且不适合变化剧烈的谱线。小波变换运算量较大，不适合处理海量数据，本系统采用中值滤波法[7]对光谱数据进行归一化。中值滤波是一种典型的非线性处理技术，方法简单易行，是一种比较实用的方法。中值滤波要求设置一个窗口，将其移动遍历各样本，用窗口内各原始值的中值代替窗口中心点的值，产生出比较平滑的输出图像。

设 $X=\{x_1, x_2, x_3, \cdots, x_n\}$ 为一条待处理的天体光谱数据，v 为中值滤波窗口半径，则在波长 i 处的流量对应的窗口中值，为，

$$y_i = Median(x_{i-v}, \cdots, x_i, \cdots, x_{i+v})$$

2.2　聚类

聚类分析输入的是一组未分类记录，聚类分析就是通过分析数据库中的记录数据，根据一定的分类规则，合理地划分记录集合，确定每个记录所在类别。简单地说，聚类是基于整个数据集内部存在若干“分组”为出发点而产生的一种数据描述，分组后使得每个子集中的点具有高度的内在相似性[8]。天体光谱离群数据挖掘系统采用的聚类方法分为两个步骤：(1)利用距离矩阵确定待处理数据集的聚类中心点；(2) 利用基于距离的方法将待处理数据集聚类。

设 $R=\{x_1, x_2, x_3, \cdots, x_n\}$ 为天体光谱数据集，其中每条光谱数据 x_i 有 m 个流量值，$d(x_i, x_j)=\sum_{k=1}^{m}(x_{ik}-x_{jk})^2$ 为 x_i 与 x_j 之间的距离，记为 d_{ij}。由 d_{ij} 的距离公式可以计算出这 n 条光谱数据项之间的 $n\times n$ 距离矩阵 $\boldsymbol{D}$，选择其中产生的最小值的数据项合并，合并方式采用向量法，使结果生成一个新点。删除以上两点，将新点插入数据集再次执行以上运算，直到产生希望得到的簇中心点个数。将结果记录下来，作为聚类的参照中心点。

确定聚类中心点需要找到距离矩阵各元素的最小值，这里采用一趟冒泡排序法的方法，接着将对应两个样本合并生成一个新对象，将新对象插入数据集同时删除以上两对象，重复以上工作直到数据集中只有 k 个对象，即为 k 个聚类的中心点。然后，以产生的 k 个点作为聚类中心点聚类，按最近分配原则把数据集中所有对象 O_i 分配到以 k_i 为中心的簇中。

2.3　离群数据挖掘方法

特定环境下对离群数据的定义标准是不同的，即使同一环境下根据不同的要求对离群数据的定义也有差异，要求发现的离群数据范围也不同。为了能够根据不同要求发现离群数据，所采用的挖掘方法应该能通过对一定约束条件的修改得到不同的挖掘结果。

本系统通过引入了距离支持度，以此确定样本与对应中心点的距离的上限阈值，如果样本到中心点的距离大于此值与平均距离之积，则认为该样本为离群数据。引入了距离支持度之后，必须对其给出一个确定的有效范围。一般来说离群数据是明显偏离其他数据、不满足数据的一般模式或行为、并与存在的其他数据不一致的数据。如果以距离为度量标准，则离群数据和中心点的距离应该大于数据集中样本间的平均距离，在此以平均距离作为加权平均距离的最小值，即距离支持度最小值为 1，在小于该值条件下发现的离群数据没有实际意义。聚类之初构造的距离矩阵最大值为数据集中属性差异最大的两元素之间的距离，任何离群数据与其他数据之间的距离不可能超过该值，因此，可以定义距离支持度最大值为该值与平均距离的比值。

实际应用中，在输出离群数据时要根据其可能性对其进行排序，虽然距离支持度变小时，输出离群数据会增加，但排在底部的点并不一定是真的离群数据。

2.4 可视化

可视化技术是帮助人们表示数据或挖掘数据隐含信息的手段，目的是辅助人们得出某种结论性观点。从常识性的认知角度而言，现实世界是一个三维空间，使用计算机将现实世界表达成三维模型从而更加直观逼真，因为三维的表达不再以符号化为主，而是以对现实世界的仿真手段为主。

天体光谱离群数据挖掘系统中首先采用二维谱线表示离群光谱的原始特征，但是光谱谱线虽然比较精确但是不够直观和形象，无法表示光谱之间的相对关系，因此该系统对光谱数据进行了三维可视化显示。由于天体光谱数据是高维数据，要实现三维可视化显示必须对数据降维。系统采用了主分量分析法 PCA[1]。实际操作中用聚类中心点代表聚类，将聚类中心点和运算得到的离群数据根据相对距离，放置在一个立方体中，加以标识。该系统实现了立方体的多角度旋转，可以从不同角度观察立方体，清楚地显示出了数据之间的相对关系。

3 系统运行结果

以国家天文台提供的 SDSS 恒星光谱数据，系统进行了试运行。数据预处理中，选择归一化窗口大小为 7，采用中值滤波对数据归一化，对于每一条纪录的前 3 个属性和最后三个属性由于没有窗口中值与之对应，系统中统一用“1”替换原值，由于光谱数据是高维数据，这部分的影响可以忽略，聚类中心点个数选择为 7，距离支持度为 350%。图 3 为聚类结果，显示了光谱数据的 ID 号、所属类别以及和该类中心点的距离，图 4 为离群数据挖掘结果，同样给出了光谱数据 ID 号、所属类别以及和该类中心点的距离，图 5 是一条离群光谱图，图 6 为离群数据的三维可视化输出。

挖掘得到的 7 条离群光谱数据，经分析认证，其中 5 条非恒星数据，两条为 O 型星，表面温度高达 55 500 K℃和 57 500 K℃，接近恒星最高表面温度 60 000 K℃，说明知识发现的过程是成功的。

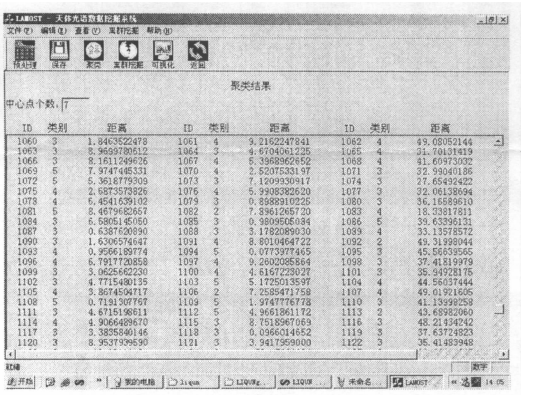

Fig. 3 Clustering result

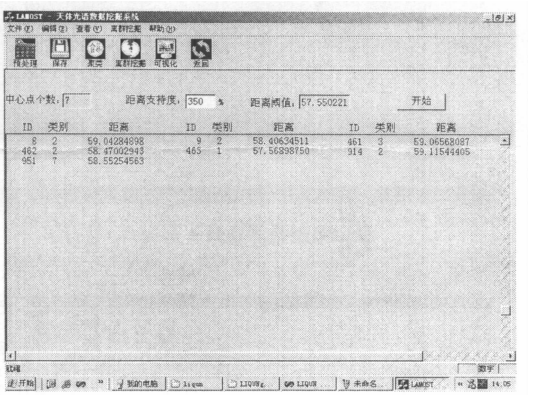

Fig. 4 Result of outlier mining

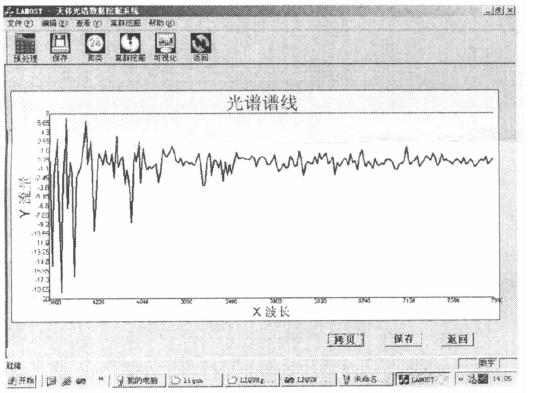

Fig. 5 Outline spectrum

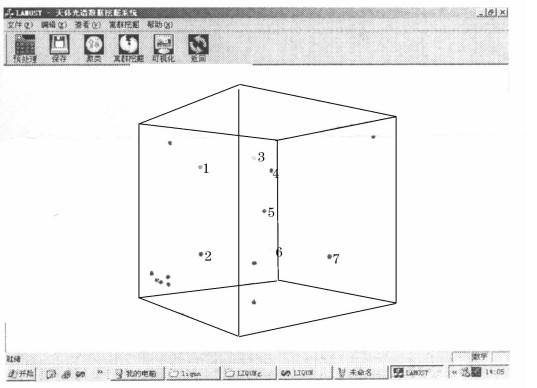

Fig. 6 Three-dimensional visualization of outliers

4　结束语

本文以 VC++和 Oracle9i 为开发工具，设计与实现了天体光谱离群数据挖掘系统，并对预处理技术、聚类、离群数据挖掘、可视化等关键技术进行了描述。但由于人类对天体认识很有限，如何正确确定天体光谱数据聚类数以及光谱数据归一化参数，将是进一步提高挖掘结果正确性的关键，也是下一步研究工作的重点。

参　考　文　献

[1] ZHAO Rui-zhen, HU Zhan-yi, ZHAO Yong-heng(赵瑞珍，胡占义，赵永恒). Spectroscopy and Spectral Analysis(光谱学与光谱分析), 2005, 25(1): 153.

[2] QIN Dong-mei, HU Zhan-yi, ZHAO Yong-heng(覃冬梅，胡占义，赵永恒). Spectroscopy and Spectral Analysis(光谱学与光谱分析), 2003, 23(1): 182.

[3] LIU Rong, LIU San-yang, ZHAO Rui-zhen(刘　蓉，刘三阳，赵瑞珍). Spectroscopy and Spectral Analysis(光谱学与光谱分析), 2006, 26(3): 583.

[4] QIU Bo, HU Zhan-yi, ZHAO Yong-heng (邱　波，胡占义，赵永恒). Spectroscopy and Spectral Analysis(光谱学与光谱分析), 2002, 22(3): 523.

[5] Barnett V, Lewis T. Outliers in Statistical Data. New York: John Wiley & Sons, 1994.

[6] Luis Malagon-Borja, Olac Fuentes. An Object Detection System Using Image Reconstruction with PCA, The 2nd Canadian Conference on Computer and Robot Vision (CRV'05), 2.

[7] HUANG Xi-tao(黄熙涛). Two-dimensional Digital Signal Processing Ⅱ: Transforms and Median Filters (二维数字信号处理Ⅱ：变换与中值滤波器). Beijing: Science Technology Press(北京：科学技术出版社), 1985.

[8] Karypis G, Han E H, Kumar V. IEEE Computer, 1999, 32(8): 68.

A Study on the Outlier Mining System for LAMOST Spectra

ZHANG Ji-fu[1,2], CAI Jiang-hui[1]

1. School of Computer Science and Technology, Taiyuan University of Science and Technology, Taiyuan　030024, China

2. National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing　100080, China

Abstract　To find unknown celestial bodies is one of main goals in mankind's universe exploration, and outlier mining is a kind of effective way of finding unknown celestial bodies from mass spectrum data. In the present work, using VC++ and Oracle9i as development tools, an outlier mining system for star spectra is designed and realized, and its software architecture and function modules are outlined. At the same time, the system's key components such as star spectrum data preprocessing based on median filters, clustering of star spectrum data based on distance, outlier mining of star spectrum data based on distance support and three-dimensional visualization of star spectrum outlier based on PCA, are elaborated. The preliminary experimental results based on SDSS star spectrum data show that the system is workable for outlier mining of celestial body spectrum data, and a new kind of effective way of finding unknown and peculiar celestial body spectrum data.

Keywords　Celestial body spectrum data; Outliers; Clustering; Distance support

(Received Dec. 5, 2005; accepted Apr. 21, 2006)