第29卷,第2期
2009年2月

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

Vol. 29, No. 2, pp551-555
February, 2009

基于约束概念格的天体光谱局部离群数据挖掘系统

张继福, 张素兰, 蒋义勇

太原科技大学计算机科学与技术学院, 山西 太原 030024

摘 要 寻找特殊的、未知的天体是人类探索宇宙奥妙所追求的目标之一, 天体光谱数据挖掘是实现该目标的一种有效方法。约束概念格是一种新的概念格结构, 具有构造效率高、提取知识针对性和实用性强等特点。针对天体光谱数据在特征子空间中的局部偏离, 采用 VC++6.0 和 Oracle 9i 作为开发工具, 设计与实现了基于约束概念格的天体光谱局部离群数据挖掘系统, 并对软件模块功能和体系结构, 以及天体光谱数据预处理、约束概念格构造方法、基于链表结构的概念格构造、局部离群数据挖掘方法等关键技术进行了详细描述。运行结果表明, 该系统实现天体光谱数据局部离群数据挖掘是可行的、有价值的, 从而为寻找未知的、特殊的天体提供了一种新途径。

关键词 天体光谱; 局部离群数据; 约束概念格; 稠密子空间; 稀疏度系数

中图分类号: TP311 **文献标识码**: A **DOI**: 10.3964/j.issn.1000-0593(2009)02-0551-05

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。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遍及天文学多个层次。预计每个观测夜晚将收集 2 万~4 万条光谱的数据, LAMOST 所观测到的光谱数据容量可达 4 TB[1]。利用传统人工处理数据方式将无法满足实际需求, 急需研究以计算机为主的全自动分析技术。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、特殊的天体。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、特殊的天体是数据挖掘值得研究和探索的新应用领域。

目前, 天体光谱数据分析与处理主要集中在分类和识别方面, 天文学界研究较多的是恒星光谱的分类, 具有代表性的成果有, Autoclass 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的光谱类型和谱线; Gulati 等人首先采用两层 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 网络平均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; 许馨等将核技巧与 Fisher 判别分析结合起来, 提出了基于广义判别分析方法对恒星、星系和类星体的光谱进行分类; 杨金福等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, 提出了一种基于核技巧的覆盖算法; 张继福等研究开发了一种天体光谱离群数据挖掘系统; 刘中田等提出了基于小波特征的 M 型星自动识别方法等[1-5]。

传统的离群数据挖掘方法, 大多数是从全局的角度去分析数据, 较难发现某些光谱特征线存在的局部偏离。约束概念格是文献[6]提出的一种新的概念格结构, 具有构造效率高、提取知识针对性和实用性强等特点。针对特征子空间中的天体光谱局部离群数据, 采用 VC++6.0 和 Oracle 9i 作为开发工具, 设计与实现了基于约束概念格的天体光谱数据挖局部离群数据挖掘系统, 并对软件模块功能和体系结构, 以及关键技术进行了详细描述。运行结果表明, 该系统实现天体光谱数据局部离群数据挖掘是可行的、有价值的。

1 基本概念

1.1 一般概念格

概念格是由德国的 Wille 教授, 在 20 世纪 80 年代初提出的一种有效的形式化数据分析工具[7]。概念格中的每个结点是一个形式概念, 由内涵(属性集)和外延(拥有该属性集的实体集)两部分组成, 其结构及其相应的哈希图形式, 反映了一种概念层次结构, 本质上体现了实体和属性之间的关

收稿日期: 2007-11-08, **修订日期**: 2008-02-12

基金项目: 国家自然科学基金项目(60773014)资助

作者简介: 张继福, 1963 年生, 太原科技大学计算机科学与技术学院教授 e-mail: jifuzh@sina.com

系，概念内涵和外延的统一，生动而简洁地表明了概念之间的泛化和特化关系，已成为一种非常有效的数据分析和知识提取工具。

定义 1 形式背景是一个三元组 $K = (G, M, I)$，G 中的元素称为对象，M 中的元素称为属性，$I \subseteq G \times M$。如果 $g \in G$ 和 $m \in M$ 在关系 I 中，则表示为 $(g, m) \in I$，即对象 g 具有属性 m。

定义 2 设 $K = (G, M, I)$ 为一个形式背景，形式概念 $J = (A, B)$ 是满足如下两个条件的一个序偶，其中，$A \subseteq G$，$B \subseteq M$，A 被称为 J 的外延(extent)，B 被称为 J 的内涵(intent)。

(1) $A = B' = \{a \in G \mid \forall b \in B, a\ I\ b\}$

(2) $B = A' = \{b \in M \mid \forall a \in A, a\ I\ b\}$

定义 3 一个形式背景 K 中所有形式概念之间的偏序关系表示为 $(A_1, B_1) \leq (A_2, B_2) \Leftrightarrow A_1 \subseteq A_2 \Leftrightarrow B_2 \subseteq B_1$，由形式背景 K 中的所有概念及概念之间的偏序关系构成了一个完全格，称为概念格。

1.2 约束概念格

在概念格的构造过程中，概念内涵所包含的属性并非都是用户感兴趣的，同时一些概念内涵所包含的属性在实际应用中并无意义。因此，可以根据用户对数据集的兴趣、了解、认识等作背景知识，指导概念格构造过程，从而提高了概念格的构造效率，提取的知识更具有针对性和实用性。采用谓词逻辑表示知识时，首先定义描述背景知识的谓词，并指出每个谓词的确切含义，然后再用连接词把有关的谓词连接起来，形成一个谓词公式以表达一条完整的背景知识。形式背景是一个二维表，可表示为一个 n 元有序组的集合，一个集合可用一个特性谓词刻画，故一个 n 元有序组的集合可用一个 n 元特性谓词刻画[6]。

定义 4 约束概念格的每一个形式概念为 $h = ((A, B), P)$，其中 P 为背景知识，且 $P((A, B)) = .T.$，$A \in \rho(G)$ 称为 h 的外延，$B \in \rho(M)$ 称为 h 的内涵，且 A，B 同时满足以下两个条件

(1) $A = B' = \{A \in G \mid \forall B \in M, A\ I\ B\}$

(2) $B = A' = \{B \in M \mid \forall A \in G, A\ I\ B\}$

称具有这种结构的概念格称之为约束概念格(constrained concept lattice)，表示为 $\langle L(G, M, I, P), \leq \rangle$，其中：$L(G, M, I, P)$ 表示满足约束条件的概念(节点)集合，$\leq$ 表示为满足约束条件概念之间的偏序关系。满足上面两个条件且 $P((A, B)) = .T.$ 的序偶 (A, B) 均属于 $L(G, M, I, P)$。

1.3 稀疏度系数与离群数据

为了度量子空间内数据的偏移程度，假设一个高维数据集中有 D 条记录，每维均离散化为 θ 个区间，且各记录间是相互独立的。从中取出 K 个属性构造一个 K 维立方体，则这 N 条记录以概率 $(1/\theta)^k$ 按柏努利概率随机分布在立方体中，每个区域内包含的平均记录数为它的数学期望 $N \times (1/\theta)^k$。文献[8]定义了如下稀疏度系数 $S(D)$，度量子空间内数据的偏移程度。

$$S(D) = (n(D) - N \times f^k) / \sqrt{N \times f^k \times (1 - f^k)}$$

其中 $f = 1/\theta$，$n(D)$ 为分布在子空间 D 内的数据(记录)个数。稀疏度系数 $S(D)$ 为负值，表明 D 中的数据个数低于期望值。$S(D)$ 的值越小，D 中包含的数据越稀疏。

在稀疏子空间中，稀疏度系数仅反映了子空间中包含的数据对象个数远小于期望值，但数据对象个数远小于期望值，可能是数据对象在更低维子空间上的过度稀疏造成的，稀疏度系数并不能正确反映稀疏子空间上的数据偏离程度。因此仅采用 $S(D)$ 来判断稀疏子空间的方法，不能保证结果的准确性[9]。数学期望表示了子空间中对象的平均个数，引入一个用户设置的系数，采用它们的乘积来度量子空间的稠密程度。

定义 5 对于一个任意的数据集，其属性集为 M，对象集为 G，且每维均离散化为 Φ 个区间，DENSE 为用户设置的稠密度系数，$\forall$由约简属性集 $P(P \subseteq M)$ 构成的约简子空间 D，且其包含的对象集为 $A(A \subseteq G)$，若 $|A| \geq \text{DENSE} \times |G| \times (1/\Phi)^{|D|}$，则称 D 为稠密子空间。

定义 6 对于一个任意的数据集，其属性集为 M，对象集为 G，$\forall$由约简属性集 $P(P \subseteq M)$ 构成的稀疏子空间 D，且其包含的对象集为 $A(A \subseteq G)$，若 $\forall$由约简属性集 $P_1(P_1 \subset P)$ 构成的约简子空间 D_1，均为稠密子空间，则称 D 为离群子空间，A 中的数据对象称之为 D 中的局部离群数据。

2 系统功能与体系结构

基于约束概念格的天体光谱局部离群数据挖掘系统，主要包括，天体光谱数据预处理、概念格构造、离群数据挖掘等功能模块，如图 1 所示，其体系结构如图 2 所示。

3 关键技术

3.1 数据预处理

在流量离散化处理中，不仅要描述天体光谱波长处的流量强度和峰宽，同时还应描述波的形状，即：吸收线还是发射线，故对于光谱不仅需考虑波长处的强度和峰宽两个因素，而且还需要考虑波的形状。将刻画和描述天体光谱数据的 44 条特征线的高度划分为 6 个区间，吸收强、吸收一般、

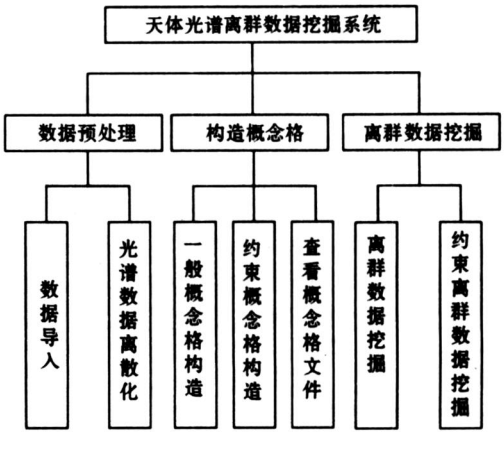

Fig 1 Function modules

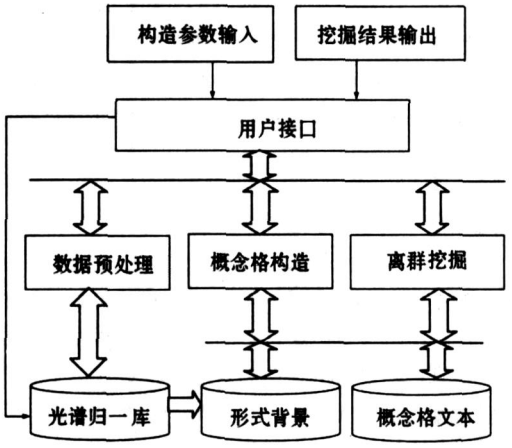

Fig. 2　Software architectural structure

吸收弱、发射强、发射一般、发射弱，特征线宽度划分为 2 个区间：窄和宽，因此根据特征线的高度和宽度的 12 种组合，对光谱特征线离散化为 13 种值。

3.2　约束概念格的构造方法

根据文献[6]，约束概念格可按如下方法来构造。

(1) 当用户关心的属性集为 X，$X \vee Y$ 或者 $X \wedge Y$，则概念格结点的内涵只能为含有 X，Y，或者 XY 的属性集。由用户关心的属性集形成的概念格为原概念格的子格，那么在概念格构造过程中，只需要对含有关心的属性的对象进行概念格的渐进式构造。

(2) 当用户不关心的属性集为 X，$X \vee Y$ 或者 $X \wedge Y$，则概念格结点的内涵由不含有 X，Y，或者 XY 的属性集，可用如下方法构造，即将含有不关心属性的对象做上删除标志，不关心的属性值记为空值；渐进式生成概念格时，如果结点为空时，先生成不含有删除标志的对象的格结点；然后求解其与前面有删除标志对象结点的关系(交集)，由交的结果决定是否生成新结点。若交集不为空时，则生成新结点，否则不做任何处理；再求解带删除标志对象结点间的关系(交集)，由交集的结果决定是否生成新结点。若交集不为空时，则生成新结点，否则不做任何处理。

3.3　基于链表结构的概念格构造

在渐进式构造过程中，若采用顺序结构(如 C++ 中的 vector 类型)存储格节点，由于生成的格节点规模比较大，且很难预先估算出最终所生成的格节点数，存储格节点所需要的内存空间超过了预先分配的连续内存空间，则需要扩充当前所分配的内存空间。若内存资源无法提供足够且与之连续的内存空间，编译器则会另开辟一块满足需要的连续内存空间，并将原来的数据拷贝到新内存空间中去，这一过程降低了程序的时间效率。如果预先分配足够大的连续内存空间以存储格节点，内存资源的浪费较大。将格节点通过链表结构存储的最大优点是格节点的物理位置可以是连续的，也可以是不连续的，避免了重新分配内存空间并拷贝数据，以及内存资源的浪费[10]。

3.4　局部离群数据挖掘

针对低维子空间中的偏移数据，文献[9]提出了一种基于约束概念格的离群数据挖掘算法 RCLOM，将约束概念格中的每个概念节点看为子空间，并计算概念节点的内涵缩减的稀疏度。若某个 K 维内涵缩减的稀疏度小于稀疏度系数阈值，考察该内涵缩减的 $K-1$ 维真子集，并判断由这些真子集构成的子空间是否为稠密的。如果这些子空间都是稠密的，则外延中所包含的对象被看作为局部离群数据。

3.5　主要功能模块的实现技术

ADO 组件提供了 VC++ 和统一数据访问方式 OLE DB 的一个中间层，具有易于使用、高速度及较低的内存占用等优势。STL 是惠普实验室开发的标准模板库，提供了 list 类型实现了链表结构的构造，查找效率很高。概念格构造主要采用 STL 技术实现。格节点通过链表结构组织，STL 提供了 list 类型实现了链表结构的构造。在判断交内涵在当前概念格节点中是否存在的过程中，采用 STL 提供 set 类型来实现。将格节点的内涵放入 set 类型的一个变量数组中，通过 set 中的 find 成员函数判断该交内涵是否已经存在。find 成员函数通过二分法进行查找，查找效率很高。离群数据挖掘主要采用 STL 和 ADO 技术实现。在求某个格节点与其父节点的差集时，采用 STL 提供的 set_difference 函数实现。在判断一个 K 维稀疏子空间的所有 $K-1$ 维子空间是否为稠密子空间的过程中，采用了 ADO 技术。

4　系统运行结果及分析

在 Pentium Ⅲ1.0G CPU，256M 内存，WindowsXP 操

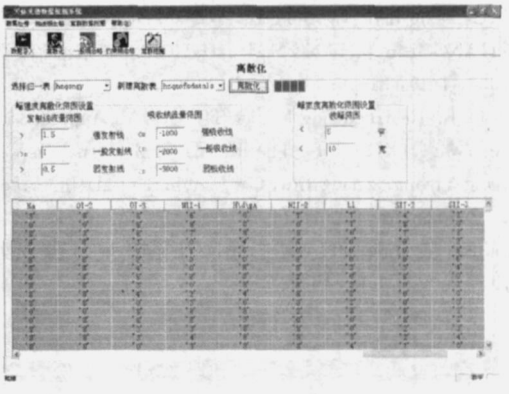

Fig. 3　Discretion of celestial body spectrum

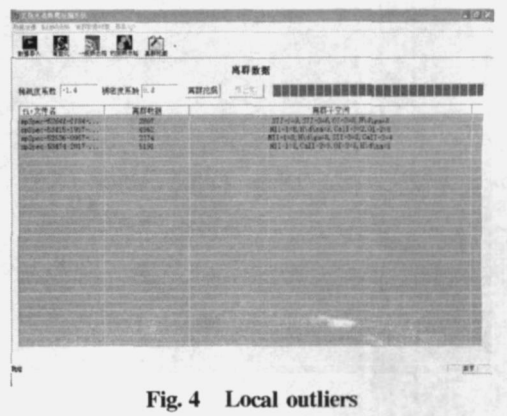

Fig. 4　Local outliers

作系统,DBMS 为 ORACLE9i,用 Visual C++6.0,设计与实现了基于概念格的天体光谱局部离群数据挖掘系统。采用国家天文台提供的 5 412 条高红移类星体光谱数据,其系统运行结果如图 3 和图 4 所示。

图 3 为高红移类星体的离散化界面,根据离散化参数(特征线的高度、宽度和形状),每条特征线可按照 3.1 节中的方法进行离散化。

图 4 为稀疏度系数为 -1.4、稠密度系数为 0.8、背景知识为(O Ⅰ-2 ∧H\d\da) ∨(S Ⅱ-2 ∧Ca Ⅱ-2)时,系统挖掘出的局部离群数据,其中第一列是该离群对象相对应的光谱数据 fit 文件名,第二列为局部离群数据的编号,第三列是局部离群数据对象所在的特征子空间。根据 RCLOF 算法可知,在该稀疏度系数阈值和稠密度系数阈值下,算法只考察外延包含的对象数为 1 的概念,并挖掘出满足定义 7 且属性数为 4 的约束内涵缩减及其对应的对象集,为离群子空间和离群数据。由图 4 可以看出,挖掘出的所有离群子空间中均包含特征线 O Ⅰ(2 和 H\d\da 或特征线 S Ⅱ-2 和 Ca Ⅱ-2,即均为用户所关心的属性集。经天文学家认证,在 4 个特征线构成的子空间中,光谱数据确实存在明显的局部偏离,从而验证了利用该系统,挖掘出的天体光谱局部离群数据是可行的、有价值的。

5 结束语

离群数据挖掘是寻找特殊的、未知的天体光谱数据的一种有效方法。利用约束概念格作为天体光谱特征的局部偏子空间表示工具,采用稀疏度和稠密度系数来度量子空间中的局部离群数据,并以 VC++6.0 和 Oracle 9i 为开发工具,设计与实现了天体光谱局部离群数据挖掘系统。系统运行结果表明,该系统为实现天体光谱局部离群数据挖掘是可行的和有价值的。

参 考 文 献

[1] 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), 2004, 24(4): 507.

[2] XU Xin, YANG Jin-fu, WU Fu-chao, et al(许 馨, 杨金福, 吴福朝, 等). Spectroscopy and Spectral Analysis(光谱学与光谱分析), 2006, 26(10): 1960.

[3] YANG Jin-fu, XU Xin, WU Fu-chao, et al(杨金福, 许 馨, 吴福朝, 等). Spectroscopy and Spectral Analysis(光谱学与光谱分析), 2007, 27(3): 602.

[4] ZHANG Ji-fu, CAI Jiang-hui(张继福, 蔡江辉). Spectroscopy and Spectral Analysis(光谱学与光谱分析), 2007, 27(3): 606.

[5] LIU Zhong-tian, LI Xiang-ru, WU Fu-chao(刘中田, 李乡儒, 吴福朝). Acta Electronica Sinica(电子学报), 2007, 35(1): 157.

[6] ZHANG Ji-fu, ZHANG Su-lan, HU Li-hua(张继福, 张素兰, 胡立华). CAAI Transactions on Intelligent Systems(智能系统学报), 2006, 1(1): 31.

[7] Wille R. Restructuring Lattice Theory: An Approach Based on Hierarchies of Concepts. in: Rival Ⅰ ed. Ordered Sets, 1982. 415.

[8] Agarwal C C, Yu P S. The International Journal on Very Large Data Bases, 2005, 14(2): 211.

[9] Jiang Yiyong, Zhang Jifu, Cai Jianghui, et al. In: Proceedings of The First International Symposium on Data, Chengdu China: Privacy, & E-Commerce, 2007. 11, 80.

[10] JIANG Yi-yong, ZHANG Ji-fu, ZHANG Su-lan(蒋义勇, 张继福, 张素兰). Computer Engineering and Applications(计算机工程与应用), 2007, 43(11): 178.

The Local Outliers Mining System of Celestial Body Spectrum Based on Constrained Concept Lattice

ZHANG Ji-fu, ZHANG Su-lan, JIANG Yi-yong
School of Computer Science and Technology, Taiyuan University of Science and Technology, Taiyuan 030024, China

Abstract It is one of the main goals in mankind's universe exploration to find unknown and particular celestial bodies. Data mining is an effective way of finding the spectrum data of unknown and particular celestial body in mass celestial body spectrum data. Constrained concept lattice, with characteristics of higher constructing efficiency, practicability and pertinency, is a new concept lattice structure. For local bias data of celestial body spectrum in characteristic subspace, the local outlier mining system of celestial body spectrum based on constrained concept lattice was designed and implemented by using VC++ 6.0 and Oracle 9i as developing tools. At the same time, its software architecture and function modules were outlined. Such key techniques for preprocessing celestial body spectrum data, the constructing method of constrained concept lattice, and the local outlier mining method were discussed in details. The running results show that the system is feasible and valuable for mining local bias data of celestial body spectrum in low dimensional characteristic subspace. Therefore, the system provides an effective means for finding the unknown and particular celestial bodies.

Keywords Celestial body spectrum; Local outliers; Constrained concept lattice; Dense subspace; Sparsity coefficient

(Received Nov. 8, 2007; accepted Feb. 12, 2008)