第28卷，第12期
2008年12月

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

Vol. 28, No. 12, pp2996-2999
December, 2008

基于约束FP树的天体光谱数据相关性分析系统研究

赵旭俊，张继福*，蔡江辉

太原科技大学计算机科学与技术学院，山西 太原 030024

摘 要 从海量天体光谱数据中挖掘光谱数据特征和物理化学性质之间内在的、隐含的相关性，是人类探索天文规律的一种有效方法。利用基于约束FP树的关联规则挖掘方法作为天体光谱数据相关性分析手段，采用VC++和Oracle9i作为开发工具，设计与实现了天体光谱数据相关性分析系统，给出了其系统的软件体系结构和模块功能，并对光谱数据预处理、背景知识表示、CFP树构造、频繁模式提取及关联规则生成等关键技术以及关键模块的实现技术，进行了详细描述。系统运行结果表明，利用关联规则来描述、分析天体光谱数据特征和物理化学性质之间存在的相关性，是可行的和有价值的，从而为寻找天体规律提供了一种有效手段。

关键词 天体光谱；数据挖掘；关联规则；FP树；约束频繁模式

中图分类号：TP311 **文献标识码**：A **DOI**：10.3964/j.issn.1000-0593(2008)12-2996-04

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(LAMOST)，是国家重大科学工程项目，也是世界上光谱获取率最高的望远镜。LAMOST具有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银河系、星系、星系团、活动星系核，直到宇宙大尺度结构等。预计每个观测夜晚将收集2万到4万条光谱的数据，LAMOST所观测到的光谱数据容量将有可能达到4TB。如此庞大的观测数据，依靠人工方法进行处理和分析，是无法完成的，高效率的多目标望远镜必须和高效率的计算方法相结合，才能自动完成LAMOST的巡天任务[1]。

当前天体光谱数据分析主要集中在分类方面，采用的主要方法：交叉相关分析与主成分分析(PCA)、人工神经网络、小波变换等。典型成果有：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；许馨等人将核技巧与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；杨金福等人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，数据挖掘技术是实现该任务的一种有效手段。

关联规则是数据挖掘中的重要研究内容之一，描述了数据集中不同属性之间的关联关系。本文以LAMOST为应用背景，采用关联规则作为天体光谱数据相关性分析方法，以VC++和Oracle9i作为开发工具，设计并实现了基于约束FP树的天体光谱数据相关性分析系统，给出了其软件体系结构和模块功能，并对系统采用的关键技术进行了详细描述。系统运行结果表明，利用关联规则来描述、分析光谱数据特征和物理化学性质之间存在的相关性，是可行的和有价值的。

1 关联规则的基本概念

给定一个交易数据库DB，$I=\{I_1, I_2, \cdots, I_m\}$为DB中$m$个不同交易项目集合，DB中每一个交易$T$就是$I$中的一组项目集，即$T \subseteq I$。

定义1：模式P定义为$I_1 \cap I_2 \cap \cdots \cap I_k$，$I_i \in I(i=1, 2, \cdots, k)$，称$P$是长度为$k$的模式。

定义2：关联规则定义为形如$A \Rightarrow B$的蕴涵式，其中$A \subseteq$

收稿日期：2007-09-26，**修订日期**：2007-12-26

基金项目：国家自然科学基金项目(60573075)，山西省自然科学基金项目(2006011041)和太原科技大学校青年基金(2007131)资助

作者简介：赵旭俊，1976年生，太原科技大学计算机科学与技术学院讲师 e-mail：zxj0226@126.com

*通讯联系人 e-mail：jifuzh@sina.com

I，$B \subseteq I$，且 $A \cap B = \varnothing$。

定义 3：模式 P 在 DB 中的支持度定义为 $\sigma(P/\mathrm{DB}) = \mathrm{DB}$ 中包含 P 的交易个数/DB 中的总交易个数。

定义 4：A 和 B 为 2 个模式且 $\{A_i\} \cap \{B_j\} = \varnothing$，其中：$A = A_1 \cap A_2 \cap \cdots \cap A_k$，$B = B_1 \cap B_2 \cap \cdots \cap B_m$，则关联规则 $A \Rightarrow B$ 在 DB 中的置信度为：$\Psi(A \Rightarrow B/\mathrm{DB}) = \sigma(A \cap B/\mathrm{DB})/\sigma(A/\mathrm{DB})$。

为了在 DB 中挖掘有效的关联规则，必须首先定义最小支持度 $\sigma_{\min}$ 和最小置信度 $\Psi_{\min}$，关联规则的挖掘就是在 DB 中寻找满足 $\sigma(A \cap B/\mathrm{DB}) \geqslant \sigma_{\min}$ 和 $\Psi(A \Rightarrow B/\mathrm{DB}) \geqslant \Psi_{\min}$ 的所有关联规则 $A \Rightarrow B$。由于 $\Psi(A \Rightarrow B/\mathrm{DB})$ 的值可由 $\sigma(A \cap B/\mathrm{DB})$ 和 $\sigma(A/\mathrm{DB})$ 的值来计算，因此挖掘关联规则 $A \Rightarrow B$ 重点在于生成 k 频繁模式集，目前大量的研究工作主要集中在 k 频繁模式集的生成问题上[5,6]，它是提高关联规则挖掘效率的关键。频繁模式生成主要有 Apriori[5] 和 FP-tree[6-8] 两类算法。

定义 5：频繁模式树(FP-tree)为满足以下 3 个条件的树型结构：①包含一个标为“null”的根节点，根节点的孩子是项前缀子树集合，该树还包含频繁项目头表；②项目前缀子树中的每一节点包含 3 个域：item-name，count，node-link，其中，item-name 记录项目名，count 记录能到达该节点的路径所表示的交易的数目，node-link 为指向 FP-tree 中具有相同的 item-name 值的下一节点，当下一个节点不存在时，node-link 为 null；③频繁项目头表的每一表项包含两个域：item-name，head of node-link，其中 head of node-link 为指向 FP-tree 中具有相同的 item-name 值的首节点的指针。

2　系统的软件体系结构及功能

基于约束 FP 树的天体光谱数据相关性分析系统，主要包括天体光谱数据预处理、背景知识表示、天体光谱数据的 CFP 树构造、天体光谱数据频繁模式提取和关联规则生成等模块，其功能模块如图 1 所示。

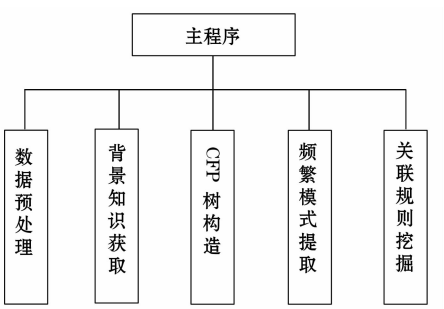

Fig. 1　Function modules

在图 2 中，给出了该系统的软件体系结构。首先通过用户接口输入数据归一化、离散化参数，对光谱数据进行归一化、离散化预处理；然后利用用户感兴趣背景知识，构造天体光谱数据的 CFP 树，并通过遍历 CFP 树，提取天体光谱数据的约束频繁模式；最后由约束频繁模式，生成刻画天体光谱数据相关性分析的关联规则。

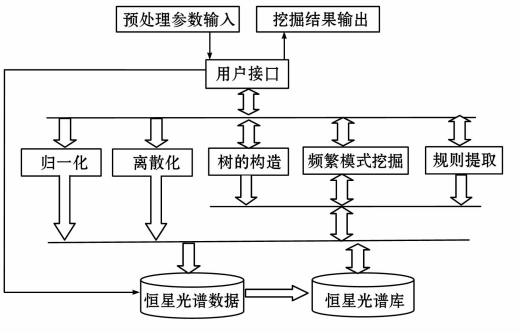

Fig. 2　Software architectural structure

3　关键技术

3.1　归一化、离散化预处理

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，而连续谱形状对天体来说是重要的分类特征，在分类中起关键作用。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 $10^{-19} \sim 10^{-3}$，很大程度上影响了运算效率，因此需要对连续谱归一化，消除数量级上的差异，本系统采用中值滤波法[9]对光谱数据进行归一化。同时为了直观描述光谱波的特征，适应数据挖掘的要求，需要对天体光谱数据进行离散化。

在恒星光谱的流量离散化中，不仅要描述波的流量强度，同时还应考虑峰的宽度。设 I 和 W 为两个特征变量，I 表示某一波长处波峰的强度，W 表示某一波长处波峰的宽度，可将光谱数据转变为以特征变量 I 和 W 表示的一系列特征数据。假设一条光谱数据中的波长表示为 $A = [A_i]$，$i = 1, 2, 3, \cdots, n$。其中 A_i 表示该光谱中的第 i 个波长，n 表示光谱数据共有 n 个波长，那么 A_i 可表示为 $A_i = \{I_i, W_i\}$。根据专家的建议，可将 I_i 等分为五种，W_i 等分为三种，因此 A_i 可被离散化为 13 种离散值之一。对于恒星光谱的物理化学性质，主要包括温度、光度、化学丰度和微湍流等的离散化。用 $S = [S_j]$ 表示各种物理化学性质组成的集合，并将 S_j 等分三种离散值，其中：$j = 1, 2, 3, \cdots, m$，S_j 表示恒星光谱的第 j 个物理化学性质。

3.2　一阶谓词逻辑与背景知识

天文学家通过长期工作经验积累，那些光谱特征对分析和识别光谱数据是重要的，或对那些光谱特征感兴趣，具有深入的认识。因此，利用天文学家积累起来的经验和兴趣作为先验信息(即背景知识)，来指导 FP 树的构造，可以有效地提高关联规则挖掘结果的针对性，降低 FP 树的复杂性，进而有效地解决 FP-tree 算法中的数据存储瓶颈问题。谓词逻辑是一种形式语言系统，它用逻辑方法研究推理的规律，适合于表示事物的状态、属性、概念等事实性的知识，可采用一阶谓词逻辑表示背景知识，指导 FP 树的构造。

定义 6：设 r 是天体光谱数据库中的关系表名变量，f 是表示关系表到属性的映射的函词，$\sigma_{\min}$ 是最小支持度($0 \leqslant \sigma_{\min}$

$\leqslant 1$)，则背景知识 G 可由如下 3 组谓词公式，通过逻辑运算符组成合适公式。

1) Interesting($f(r)$)；2) support($f(r)$，$\sigma_{\min}$)→Interesting($f(r)$)；3) Interested($f(r)$)→Interesting($f(r)$)

在定义 6 中，r 是天体光谱数据库 DB 中的关系表名变量，f 是函词，表示关系表到属性的映射，$f(r)$是关系表 r 中的属性集合，也就是离散化后天体光谱的项目子集。Interesting($f(r)$)是结论谓词，描述了对包含有项目集 $f(r)$的光谱模式感兴趣。谓词 support($f(r)$，$\sigma_{\min}$)的含义是光谱项目子集 $f(r)$的支持度大于给定的最小支持度 $\sigma_{\min}$。设光谱属性集合 $f(r)$是历史挖掘中的频繁模式子集，则谓词 Interested($f(r)$)描述了以后挖掘中，对 $f(r)$也是感兴趣的。第一个谓词公式表示，用户直接给出感兴趣的天体光谱项目集；第二个谓词公式表示，在历史挖掘中，如果天体光谱项目集 $f(r)$的支持度大于 $\sigma_{\min}$，那么在以后挖掘中也是用户感兴趣的光谱项目集；第三个谓词公式表示，如果光谱项目子集 $f(r)$是历史挖掘中的频繁模式子集，那么在以后挖掘中用户对 $f(r)$也是感兴趣的。

3.3 天体光谱数据的 CFP 树构造

定义 7：设 G 为天体光谱数据的背景知识，对于任意 FP 树，如果从根节点到叶子节点的路径中所描述的任一频繁模式 P，使得 $G(P)$=True，则称 FP-Tree 为天体光谱数据的约束频繁模式树(CFP 树或约束 FP 树)。

CFP 树的构造采用两次扫描数据库来完成(CFP-Construct)：

(1) 扫描天体光谱数据库 DB 一次，收集 1-频繁模式的集合和它们的支持度，对 1-频繁模式按支持度降序排序，结果为频繁项表 L；

(2) 创建天体光谱数据的 CFP 树的根节点，以“null”标记；

(3) 对于 DB 中每个事务 T，如果 T 中不包含用户感兴趣模式，则扫描下一个事务，否则，执行(4)；

(4) 将 T 中频繁项按 L 中的次序排序为 T'，并按如下来更新 CFP 树：

① 在 CFP 树中，寻找与 T'的最长前缀匹配的路径；

② 对于该匹配路径上的节点，其计数增加 1；

③ 找出 T'中未匹配后缀，以确定最长匹配前缀中的最后一个频繁项所对应节点，作为根节点，依次在 CFP 树创建孩子节点，并置其计数值为 1。

3.4 天体光谱数据的频繁模式提取与关联规则挖掘

参照 FP 树的频繁模式提取过程[6]，通过遍历天体光谱数据 CFP 树，来实现天体光谱数据频繁模式及其支持度的提取。根据定义 4 和最小置信度 $\Psi_{\min}$，由频繁模式，生成关联规则及其置信度。

3.5 关键模块的实现技术

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。天体光谱数据的归一化、离散化过程中，需要对光谱原始数据进行扫描，由于库的庞大，扫描将耗费大量内存和时间，因此采用 ADO 技术动态连接、访问数据库。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。天体光谱数据的 CFP 树构造过程采用 STL 技术实现；CFP 树节点通过链表结构组织，索引表采用顺序结构组织，保证了处理速度和内存方面的高效性，提高了代码的可读性及系统的可维护性。

在天体光谱数据 CFP 树构造、频繁模式提取及相关性分析等功能中，本系统采用多线程技术，使得各线程共享进程的虚拟地址空间，访问进程的资源，处于并发执行状态，相应地提高了运行效率。

4 系统运行结果及分析

基于以上所述，采用 VC++和 Oracle9i 作为开发工具，在 Pentium Ⅳ-3.0G CPU，512M 内存，Windows XP 操作系统上，设计与实现了天体光谱数据相关性分析系统。采用 SDSS 恒星光谱数据，并选定了 200 个波长和恒星的 5 个物理化学性质作为属性集。

图 3 是离散化参数设置界面，描述了天体光谱数据的离散化参数设置，并选择归一化数据表，然后进行离散化，并将离散化结果保存到相应的数据表中，形成了面向关联规则挖掘的天体光谱数据库。

Fig. 3 Discretization parameter setting

图 4 是相关性分析结果界面，其中：$\sigma_{\min}$=1%，$\Psi_{\min}$=70%。最后一条关联规则是：3 870_强宽，4 090_较弱宽，4 850_较弱宽，5 250_强宽=〉温度_D，化学_2，微湍流_2，光度_2(1.800%，78.26%)，含义：如果(1)在波长为 3 870

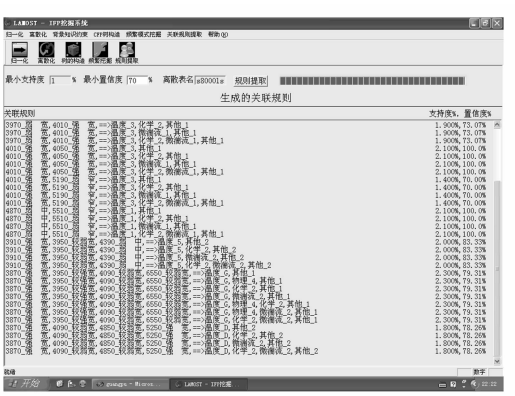

Fig. 4 Association rules mining

处有很强且很宽的峰、(2)在波长为 4 090 处有较弱且很宽的峰、(3)在波长为 4 850 处有较弱且很宽的峰、(4)在波长为 5 250处有很强且很宽的峰；那么此光谱的温度的范围为 7 500～8 300 K，化学丰度的范围为－3～－0.5，微湍流的值为 2，光度的范围为 0～1.1。该规则的支持度为 1.8%，置信度为 78.26%。将这条规则与光谱数据经验总结得出的波的特征和物理化学性质关系进行比较，发现它与 *A* 型星的特征基本类似，说明知识发现的过程是成功的。

5　结束语

本文采用关联规则作为天体光谱数据相关性分析手段，设计与实现了基于约束 FP 树的天体光谱数据相关性分析系统，并对归一化技术、离散化技术、背景知识表示、CFP 树构造、频繁模式集挖掘及关联规则提取等功能进行描述。系统运行结果表明，利用关联规则作为天体光谱数据相关性分析方法是可行的、有价值的，从而为寻找未知的天体规律提供一种有效方法。

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] Agrawal R, Imielinski T, Swami A. Mining Association Rules between Sets of Items in Large Databases. In: Proc. of 1th Int. Conf. on Management of Data, Washington DC, USA, 1993. 207.

[6] Han Jiawei, Pei Jian, Yin Yiwen, et al. Data Mining and Knowledge Discovery, 2004, 8(1): 53.

[7] Pei Jian, Wang Haixun, Liu Jian, et al. IEEE Transactions on Knowledge and Data Engineering, 2006, 18(11): 1467.

[8] Gudes Ehud, Shimony Solomon Eyal, Vanetik Natalia. IEEE Transactions on Knowledge and Data Engineering, 2006, 18(11): 1441.

[9] 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.

Research on the Interrelation Analysis System of Celestial Spectrum Data Based on Constraint FP Tree

ZHAO Xu-jun, ZHANG Ji-fu*, CAI Jiang-hui

School of Computer Science and Technology, Taiyuan University of Science & Technology, Taiyuan　030024, China

Abstract　It is an effective method of the mankind seeking after the celestial law that the inherent and unknown interrelationships between characteristics of celestial spectrum data and its physical and chemical properties are mined from the mass celestial body spectrum data. In the present paper, the interrelation analysis system of celestial body spectrum data based on constraint FP tree is designed and implemented by using the association rule based constraint FP tree as the way of analyzing celestial spectrum data, and adopting VC++ and Oracle9i as the development tools. At the same time, its software architecture and function modules are outlined. Its key techniques such as preprocessing of celestial body spectrum data, background knowledge representing, constraint FP-tree constructing, constraint frequent patterns and association rules mining etc are discussed in details. The running results show that the system is feasible and valuable for adopting association rule to describe the above interrelationships. Therefore, the interrelation analysis system of celestial body spectrum data provides an effective means for seeking after the inherent and unknown celestial law.

Keywords　Celestial body spectrum; Data mining; Association rules; FP tree; Constrain frequent pattern

* Corresponding author

(Received Sep. 26, 2007; accepted Dec. 26, 2007)