第 20 卷 第 4 期 天 文 学 进 展 Vol.20, No.4
2002 年 12 月 PROGRESS IN ASTRONOMY Dec., 2002

天文学中的数据挖掘和知识发现

张彦霞 赵永恒 崔辰州
(中国科学院国家天文台 北京 100012)

摘 要

综述了数据挖掘和知识发现在天文学中兴起的必然性及其近几年的发展状况、实现过程和具体任务，分析了当前天文数据的复杂性，介绍了天文学中数据挖掘的科学要求，系统地概括了近年来天文学中数据挖掘和知识发现领域研究的进展及其热点，并阐述了其所面临的挑战。天文学中的数据挖掘和知识发现的兴起将对天文学的发展起到巨大的推动作用，同时也在知识和技术等方面对天文学家提出了新的要求，另外，数据挖掘技术能否在虚拟天文台中成功应用，是虚拟天文台充分发挥作用的关键所在。

关 键 词 数据挖掘 — 知识发现 — 方法 — 数据分析

分 类 号 P1, N37, TP39

1 引 言

由于各种技术 (如计算机技术、互联网技术、空间观测技术等) 的飞速发展，各个领域都正面临着一场“数据爆炸”，即数据量呈指数增长。预计在未来十年里产出的数据量将超过过去所有数据量的总和。尽管目前分析和处理数据的方法及技术还远远滞后于数据量的增长，但人们已逐步意识到这些数据量的大小以及蕴含在其中的威力。

天文学也不例外，地面和空间天文台的建立、探测器像素呈摩尔规律增长、巡天技术的发展，都给天文学带来了革命性的变化：数据量通常以万亿字节 (TB)，甚至千万亿字节 (PB) 计量。更多的地面和空间天文设备、更大口径和更精密仪器的投入使用，将使天文数据进一步突飞猛长，例如哈勃空间望远镜每天大约产出 5GB 的数据量，筹建中的大口径综合巡天望远镜 (Large-Aperture Synoptic Survey) 日产数据量将高达 10TB。因此，Szalay 认为天文学正在经历着一场“数据雪崩” [1]。面对海量数据，我们将面临许多实质性的挑战，例如怎样记录、加工原始数据；怎样通过现代计算机硬件和网络系统存储、合并、获取数据；怎样快速有效地探索及分析数据并将这些数据可视化。在这种形势下，各国都在酝酿筹建全球性的虚拟天文台，而数据挖掘和知识发现正是虚拟天文台成功运行的重要因素 [2,3]。

2002-04-28 收到 2002-06-24 收到修改稿

随着计算机技术、数据库技术、统计学、数学、机器学习等方面近几十年的长足进步，数据挖掘和知识发现从中分流并发展成为一门新型学科。知识发现是对数据抽取和精化取得新知识的模式，是数据库研究中的一个很有价值的新领域。它融合了数据库技术、人工智能、机器学习、神经网络、统计学、模式识别、知识库系统、知识获取、信息检索、高性能计算、数据可视化等多个领域的理论和技术。目前，由于具有统一的组织结构、一体化的查询语言，关系之间及属性之间具有平等性等特点，关系型数据库被广泛应用。因此，数据库知识发现 (Knowledge Discovery in Database ， KDD) 的研究非常活跃。KDD 最早由 Fayyad 于 1989 年提出，并定义它是从数据库中识别出有效的、新颖的、潜在有用的，以及最终可理解的模式的非平凡过程。有关这方面的课题和方法可参看 Fayyad 等人 [4] 的文章。由于知识发现是一门受到来自各种不同领域的研究者关注的交叉学科，因此有许多不同的术语名称。除 KDD 外，主要还有以下若干称法："数据挖掘" (data mining) 、"信息抽取" (information extraction) 、"信息发现" (information discovery) 、"智能数据分析" (intelligent data analysis) 、"探索式数据分析" (exploratory data analysis) 、"信息收获" (information harvesting) 和"数据考古" (data archeology) 等等。其中常用的术语是"数据挖掘"和"知识发现"。相对而言，"数据挖掘"主要流行于统计领域、数据分析、数据库和管理信息领域；而"知识发现"则主要流行于人工智能和机器学习领域。

具体到天文学中，数据挖掘和知识发现 (Astronomical Data Mining and Knowledge Discovery From Astronomical Database) 是指从天文数据中提取信息和发现知识，更具体地说，就是从海量数据中发现稀有的天体与现象，或者发现以前未知种类的天体和新天文现象。近年来，这方面的研究已成为天文数据研究领域的热点。

2　数据挖掘和知识发现

2.1　数据挖掘和知识发现的过程

数据挖掘和知识发现过程可粗略地分为三步：数据准备、数据挖掘以及结果的解释评估。

数据准备又可分为三个子步骤：数据选取、数据预处理和数据变换。数据选取的目的是确定发现任务的操作对象，即目标数据。具体地说，是根据用户的需要从原始数据库中抽取一组目标数据，以供用户使用；数据预处理一般具有消除噪声、推导计算缺值数据、消除重复记录、完成数据类型转换等功能。当数据挖掘的对象是数据仓库时，一般来说，数据预处理已在生成数据仓库时完成了；数据转换的主要目的是消减数据维数 (或降维) ，即从初始特征中找出真正有用的特征以去掉那些无关的特征或变量。

数据挖掘时首先要确定挖掘的任务或目的，随后再选择挖掘算法。同样的任务可用不同的算法实现，选择算法时须考虑两个因素：①不同的数据有不同的特点，因此要选择与之匹配的算法；②根据用户或实际运行系统的要求，如有的用户希望获得描述性的、容易理解的知识，而有的用户则希望获得预测准确度尽可能高的预测性知识。数据挖掘算法是知识发现的核心，若要获得好的挖掘效果，必须充分理解各种挖掘算法的要求或前提假设条件。

当数据挖掘结束后，需要对其结果进行解释和评价。发现的模式经过评价，可能存在冗余或无关的模式，这就需要剔除；也可能发现的模式不符合要求，这就需要重新进行挖掘，如重

新选取数据、采用新的数据变换方法、设定新的数据挖掘参数值，甚至换一种挖掘算法。为了使结果更易理解，可应用可视化技术将结果转换为人们易懂的表示形式。

在使用数据挖掘和知识发现时应注意：

(1) 数据挖掘仅仅是整个过程的一个步骤，数据挖掘的质量完全依赖于采用的数据挖掘技术和选用的数据的数量与质量。如果选择的数据不当或对数据进行了不适当的转换，则挖掘的效果就不会好。另外，所选用的数据样本要完备，否则得到的规则的推广性会很差。

(2) 整个挖掘过程是一个不断反馈的过程。如在挖掘过程中发现选择的数据不太好，或采用的挖掘技术不当都不会产生预期的效果，这就需要重复前面的过程，甚至从头开始。

(3) 可视化技术在数据挖掘的各个阶段都扮演着重要角色。在数据准备阶段，用散点图、直方图等可视化技术可以对数据有一个初步了解，以便更好地选择数据，如去掉一些极大和极小的离群数据；在数据挖掘阶段，可以通过可视化技术对挖掘过程有一个直观的了解，控制挖掘过程；在表示结果阶段，可视化技术可以帮助人们更好地理解数据挖掘的最终结果，并对这些结果做出合理的理论解释。天文学中的可视化工具有 XGobi 、 ExplorN 、 ViSta 、 CViz 、 IVEE 和其它一些软件包，它们可提供二维或多维数据浏览、平行坐标图等服务 [5] 。

2.2 数据挖掘和知识发现的任务

数据挖掘时首先要确定挖掘的任务或目的，如数据的总结、分类、聚类、关联规则发现或序列模式发现等。

分类在数据挖掘中占有重要地位。分类的目的是提出一个分类函数或分类模型 (也常称作分类器)，该模型能把数据库中的数据项映射到给定类别中的某一个。分类和回归都可用于预测。分类器的构造方法有统计方法、机器学习方法、神经网络方法等等。统计方法包括贝叶斯法和非参数法 (近邻学习或基于范例的学习)；机器学习包括决策树法和规则归纳法；神经网络方法主要是前向神经网络的反向传播算法 (Backpropagation Algorithm， BP 算法) 和 Kohonen 学习矢量量化方法 (Learning Vector Quantization, LVQ)。此外，最近又推出了一种新的分类器构造方法 —— 粗糙集 (Rough Set)。

聚类是根据不同特征，将数据划分为不同的数据类。其原理是使得属于同一类别的个体之间的距离尽可能的小，而不同类别的个体间的距离尽可能的大。聚类方法包括统计方法、机器学习方法、神经网络方法和面向数据库的方法。在统计方法中聚类亦称聚类分析，是多元数据分析的三大方法之一 (另两种是回归分析和判别分析)。在机器学习中聚类称作无监督或无教师归纳。与分类学习相比，分类的对象是有类别标识的，而聚类是无标识的。

相关性分析的目的是发现特征之间或数据之间的相互依赖关系。强的依赖关系反映的是固有的结构而不是新的或感兴趣的事物，这些知识可被其它模式抽取算法使用。常用的技术有回归分析、关联规则等。

偏差分析包括分类中的反常实例、例外模式、观测结果与期望值的偏离以及量值随时间的变化等，其基本思想是发现观测结果与参照量之间的有意义的差别。通过发现离群数据 (outliers)，可以发现一些不同寻常或奇异的天体，如褐矮星和高红移类星体的发现。

3 天文数据的特点和复杂性及其数据挖掘的科学要求

天文数据的复杂性特征很大程度上是由其特点决定的，并已成为天文数据挖掘研究中首要解决的问题。

3.1　天文数据的特点

天文数据可以从天文观测、数值模拟等途径获得。其形态有数字、符号、图形、图像等；组织方式也各不相同，有结构、半结构和非结构数据。由于空间属性的存在，天体才有了空间位置和距离的概念，而且相邻天体之间存在一定的相互作用，天文数据之间关系的类型也由此更为复杂，从而使天文数据与其它类型数据的挖掘方法存在着差异。

3.2　天文数据的复杂性

近年来天文观测技术的飞速发展，使天文数据越来越复杂化，具体表现在：

(1) 天文数据的海量

天文数据将以 TB 甚至 PB 量级计量，如此大的数据量常使一些方法因算法难度或计算量过大而无法得以实施，因而知识发现的任务之一就是要创建新的算法策略，开发新的高效算法以克服由海量数据造成的技术困难。

(2) 天文数据属性之间的非线性关系

天文数据属性之间的非线性关系是整个天文学领域复杂性的重要标志，其中蕴含着领域内部作用的复杂机制，因而被视为天文数据知识发现的主要任务之一。

(3) 天文数据的高维性

多波段性是指天文数据在不同观测波段上所遵循的规律以及体现出的特征不尽相同。这是天文数据复杂性的又一表现形式。天文数据的属性增加极为迅速，例如由于空间技术的日新月异，覆盖的波段的数目由几个增加到几十个甚至上百个，如何从几十甚至几百维空间中提取信息、发现知识则成为研究中的又一重要任务。

(4) 天文数据的缺值

缺值现象起源于某种不可抗拒的外力 (如仪器的灵敏度低、天气恶化等，一些天体在一个或多个波段探测不到，从而缺乏该波段的测量属性) 而使数据无法获得或丢失，如何对丢失数据进行恢复并估计数据的固有分布参量，成为解决数据复杂性的难点之一。

天文数据所表现出的上述复杂性特征为相应的数据挖掘和知识发现研究提出了更高的要求，并成为推动其发展的强大动力。

3.3　天文学中的数据挖掘的科学要求

数据挖掘是利用复杂的技术建立模型，然后从数据中发现模式和相关性。模型分为两类：描述性模型和预测性模型。描述性模型，即描述数据中的模式，并用以创建有意义的群或子群；预测性模型，即利用从已有的数据中推出的模型来预测未知事件。数据挖掘分为事件性数据挖掘和相关性数据挖掘。事件性数据挖掘又分为四类：①已知事件 / 已知算法：用已有的物理模型去确定数据中存在的人们感兴趣的已知现象，无论空间上还是时间上；②已知事件 / 未知算法：用模式识别或数据的聚类特性来发现已知现象中存在的新的观测相关性；③未知事件 / 已知算法：以天文现象的观测参数中存在的预期的相关性来预测数据中存在的以前未知的事件；④未知事件 / 未知算法：用临界值确定瞬时事件或独特事件，从而发现新现象。相关性数据挖掘则分为三类：①空间相关：证认在天空同一位置中的天体；②时间相关：证认发生在相同时间或相关时间内的事件或现象；③一致相关：用聚类方法证认存在于同一多维参

数空间的现象。

简而言之，天文学中的数据挖掘的科学要求有以下几种 [6]：

(1) 天体的交叉证认：以源的位置为参量，将存在不同数据库中的源联系起来，用以加深对证认源的新的天文理解。例如寻找 γ 暴对应体。

(2) 天体的交叉相关：用假定分析方法处理数据中的所有参数，分析天体在各个参数空间中的分布以及参数间是否具有相关性。例如在 HDF (Hubble Deep Field) 巡天中，通过双色图利用 U 波段的“ dropouts ”证认远距离星系；在 DPOSS (Palomar Digital Sky Survey) 和 SDSS (Sloan Digital Sky Survey) 巡天中，通过天体在双色图中远离正常恒星区的特征发现了高红移类星体。

(3) 最近邻规则证认：在多维空间中运用聚类算法证认天体或天文现象。如在 TW 长蛇座中通过天体具有相似的运动学特征、 X 射线发射特征、 Hα 线特征和 Li 丰度，发现了人们最熟悉的年轻恒星族。

(4) 系统的数据探索：在数据库中广泛地应用事件性和相关性数据挖掘技术可以偶然发现一种新天体或新类型天体。例如新一类变星的发现，在 MACHO (MAssive Compact Halo Object) 数据中发现了“ bumpers ”。

4 天文学中的数据挖掘技术

4.1 针对海量数据的算法研究

支持数据挖掘技术的三种技术是海量数据收集、强大的多处理器计算机、数据挖掘算法。因此要想提高算法效率，必须从以下三个基础做起：

(1) 建立虚拟天文台。正在建设中的虚拟天文台将把由空间与地面观测设备获得的多波段巡天的海量数据有机地结合起来，同时将提供利用这些数据资源进行科学研究所必需的各种计算机及网络方面的软硬件资源，从而使天文学家可以获得高数量高质量的数据，通过分析探索找到隐含在这些数据中的一些至今尚未解决的问题的答案 [2,3]。

(2) 改变算法运行的策略。其主要方式为采用并行运算环境，实施并行算法。如在大型数据库中实施决策树分类、空间聚类以及关联规则发现等算法，由此可大幅度提高计算效率。至于提高数据库查询语言的效率，则要求大型分布式的数据库不仅能实现数据的一体化存储，解决数据的索引、组织和分布管理问题，还需要有一体化的查询语言作为操作的接口，只有这样才能实现对数据库的快速查询，如目前流行的 SXQL [7] 和 XML 语言 [8,9]。

(3) 发展新的有效算法或对原有算法的结构进行改进以及多种算法的交叉和混合使用，减小运算的复杂度。 Auton 实验室小组提出了一种称为 CSS (Cached Sufficient Statistics) 的方法，它能自动地从大量的数据中挖掘和发现新知识 [10]。 Nichol 等人 [11] 在计算机科学和统计学基础上开发了一些高效而快速的聚类算法，用以从多维的天文数据库中发现星系团，并将错误发现率 (False-Discovery Rate, FDR) 这一概念引入天文学。关于错误发现率在天文学中的应用可参考 Miller 等人的文章 [12,13]。 Komarek 和 Moore [14] 将静态的 AD 树从结构上调整为动态的 AD 树，从而克服了静态的三个弱点。 AD 树是一种从数据库中快速计数和快速学习相关规则的方法 [15]。 Pelleg 和 Moore [16] 为了有效地估计数据中的类别数，将 K 均值扩

展为 X 均值。 Moore [17] 在多分辨率 KD 树的基础上研究出一种新方法：混合模型聚类法，减小了以 EM (Expectation Maximization) 算法为基础的聚类算法的复杂度。 Maino 等人 [18] 在独立分量分析 (Independent Component Analysis, ICA) 方法的基础上开发了一种新而快的算法 FastICA (Fast Independent Component Analysis)，用以分离天体的物理参量。

4.2　神经网络

神经网络是模仿人脑神经网络的结构和某些工作机制而建立的一种计算模型 [19]。其特点是将大量的简单计算单元连成网络来实现大规模并行计算。由于神经网络非常适合处理天文数据的非线性复杂关系，且在处理复杂问题时不需要了解网络内部所发生的结构变化，因而被广泛应用于天文数据挖掘和知识发现的研究中。它以不同的网络结构实现了空间聚类、分类、关联、回归、模式识别等多种算法。例如：自组织映射 (SOM) 具有无监督性自动提取特征，以及主分量分析、聚类、编码和特征映射等功能，有助于可视化。它将高维数据投影到二维平面上，保持拓扑映射 [20]；学习矢量量化 (LVQ) 区别于 SOM 而具有监督性，其网络结构类似 SOM，但无拓扑结构，每一个输出神经元代表一个已知的种类 [21]。

神经网络在天文学中的应用十分广泛，如星表的提取 [22]、恒星与星系的分类 [23~26]、星系形态的分类 [27,28]、恒星光谱的分类 [29~31]，在多参数空间中寻找具有预测特性的已知类型天体 (如寻找高红移类星体) 等。 Next (Neural Extractor) 是一个建立在神经网络基础上的软件包，它可以自动地从天文图像中提取星表、寻找特殊天体和进行恒星 / 星系分类 [32]。

4.3　统计方法

统计方法是从事物的外在数量上的表现去推断该事物内在可能存在的规律性。常用的方法有回归分析 (多元回归和自回归等)、判别分析 (贝叶斯判别和非参数判别等)、聚类分析 (系统聚类和动态聚类等) 以及探索性分析 (主分量分析法和相关分析法等) 等。有关天文学中的统计方法的详细评述可参看 1996 年、 1997 年 Babu 和 Feigelson [33,34] 的文章，大多数多变量方法的 Fortran 程序可参考 Murtagh 和 Heck [35] 的程序。

EM 算法是一种聚类算法 [36,37]，在天文学中常用于两种情形的密度估计：星系在红移空间的聚类；恒星在色空间的聚类。 EM 算法提供了星系在红移空间的平滑分布，准确地描述了数据库中数据量的大小范围特征，并且提供了一种证认多维色空间中的远离正常恒星的天体的方法，例如高红移类星体的证认。

主分量分析 (Principle Component Analysis, PCA) [38~40] 是一种线性分析方法，它具有非监督性，能降维去噪，常用于数据的预处理，可帮助去掉一些无关或不重要的参量。在天文学中主要用于恒星、星系和类星体的光谱分类；星系的形态分类；红移的自动确定；通过将发射线分解为几个独立量来研究发射区中发射线的变化及其结构和动力学特征；在观测基平面，即多维参数空间的一个子空间中，依据星系的形态、测光和动力学分类来研究低红移星系和高红移星系。

统计学习领域的研究热点 —— 支持向量机 (Support Vector Machine, SVM) 是一种基于统计学习理论的一般性构造学习方法，其主要思想是在高维空间内利用线性函数的对偶核，并通过内积空间的向量运算来处理线性不可分数据。支持向量机模型在学习效率、解决过度拟合问题、全局最优化等方面都表现出优于神经网络的良好性质；在解决天文数据的分类、特征识别、图像压缩等问题方面也有一定的进展。从 SVM 产生的背景和应用的效果来看，该模型

特别适合处理高维、复杂的目标识别问题，例如利用天体的多波段数据对天体进行分类 [41] 。关于支持向量机的原理的详细介绍可参考 Burges [42] 的文章。

4.4 模糊集

对于天文学中的一些不确定属性，通常采用模糊集理论加以描述。该理论的优势在于利用隶属函数来刻画属性的不确定性，用部分归属代替了归属的概率。然而，隶属函数虽然对部分确定关系进行了成功的刻画，打破了非此即彼的传统概念，但其确定仍然需要借助先验知识，从而导致了结果的多解性。目前，模糊集的思想已渗透到天文学数据挖掘和知识发现的各种方法中，如模糊聚类与分类、模糊神经网络、模糊专家系统等等。 1992 年 Spiekermann [43] 开创性地将模糊几何及相应的启发式方法引入天文学，对星系的形态进行分类。 Mahonen 和 Frantti [44] 利用模糊分类器对恒星和星系的图像进行分类。

4.5 高维数据的挖掘算法

在近期的研究中，对高维数据进行挖掘的思路一般有两条，一是将高维数据通过线性变换投影到低维空间，然后再实施其它挖掘算法；另一种就是采用适合处理高维数据的算法直接对其进行信息提取。

降维方法的主要问题在于，当维数无限增加时，由高维到低维的线性变换会掩盖数据原有的信息，而使数据呈现正态分布。这样原先在高维空间中存在明显差异或特征的类别在低维空间中会混杂在一起难以区别，因而高维空间向低维空间线性变换的关键就在于寻找合适的投影方向，将高维空间的目标特征信息尽可能忠实地投影到低维空间。

由高维向低维空间进行线性投影的方法有多种，最常见的有主分量分析 (PCA) 、 MDS (MultiDimensional Scaling) 、空间因子分析及其相应的改进方法等。针对非线性情况， Tenenbaum 等人 [45] 提出了 Isomap (Isometric feature mapping) 方法， Roweis 和 Saul [46] 提出了 LLE (Locally Linear Embedding) 方法，这两种方法都可用于非线性高维数据的处理。除此之外，降维方法还可通过使用粗糙集理论精简维数来加以实现；而支持向量机则能够应付处理数据时因维数过高而产生的复杂性。

4.6 天文学中的常见问题及其处理

在天文学中会遇到各种各样的问题，而这些问题如何处理，是摆在天文学家面前的重要课题。表 1 列举了一些天文学中经常遇到的问题及其处理方法。

然而天文学中的问题远远不止这些，例如宇宙大尺度结构和银河系结构图像及其定量分析、各种天体 (特殊种类或特殊性质的恒星或星系、活动星系核、星系团等) 完备样本的建立与研究等。

任何一种算法都有其优劣性，对这个问题适用的算法可能对另一个问题无效；或者几种算法处理问题的效果相当；或者用一种算法无法解决的问题，通过几种算法的混合使用会收到意想不到的效果。例如 Cortiglioni 等人 [47] 通过对比自组织映射、学习矢量量化，以及利用模糊分类器和 BP 神经网络 (Back Propagation Network) 的混合算法在大规模巡天中自动区分恒星与星系的效果，得出：好的分类效果依赖于算法的复杂性和可获得的训练样本 (training sample) 。 Lahav [56] 对星系光谱的压缩与分类方法进行了评述，着重介绍和对比了三种方法：主分量分析 (PCA) 、信息瓶颈 (IB) 、 Fisher 矩阵 (FM) 。 PCA 和 FM 属于线性分析，而 ICA 和 IB 属于非线性分析；与 FM 相比， PCA 和 IB 在模型上是独立的，但 IB 监督的波长群在

表 1　天文学中常遇到的问题及其处理方法

问　题	例　子	常　用　方　法
天体分类	恒星 / 星系分类 星系形态分类 恒星 / 星系 / 类星体分类	学习矢量量化 (LVQ) [47] 支持矢量机 (SVM) [41] 主分量分析 (PCA) [48] 自组织映射 (SOM) [47,49] 模糊集理论 [43,44,47] 神经网络 (NN) [23~31,50] 小波变换 [48] 决策树 [24]
图像分类	数字巡天中的恒星 / 星系分类	学习矢量量化 (LVQ) [21] 自组织映射 (SOM) [51] 模糊集理论 [52] 神经网络 (NN) [22] 最近邻规则 [53] 聚类分析 [54] 决策树 [55]
数据压缩与分类	光谱压缩和分类	主分量分析 (或 KL 变换) [56] 独立分量分析 (ICA) [57] 信息瓶颈 (IB) [56] Fisher 矩阵 (FM) [56]
重建方法	大尺度巡天中的图像重建	均方差估计 (UMV) [58] 小波分析 [59] 维纳滤波 [60] shapelet 公式 [62] 傅里叶拟合 [62] 变像素线性重建 [63] 最大熵方法 (MEM) [64] Massive Inference 方法 [64] Pixon 方法 [64]
大尺度结构分析	有关大尺度结构和微波背景辐射的大尺度巡天	独立分量分析 (ICA) [57] 最大熵方法 (MEM) [57,65] 贝叶斯分析 [60] 小波分析 [57,66] 错误发现率 (FDR) [67] N 点相关函数 [68] FastICA 方法 [18]

概念上接近 FM ； ICA 在计算上比 PCA 复杂，数据压缩效率弱于 PCA ，但可以较好地分离混合变量；与 PCA 方法相比， ICA 对位置、方向以及带通选择的特征量比较敏感。 Lasenby 等人 [64] 介绍了贝叶斯的求逆和正则化原理，尤其讨论了最大熵方法的概念基础，也讨论了一

些以贝叶斯为基础的消卷积方法，并且以天文学和非天文学中的应用为例比较了维纳滤波方法、 Massive Inference 和 Pixon 方法。至于最大熵方法在天文数据处理中的应用可参考 Starck 等人 [69] 的文章。Kim 等人 [70] 对比了三种聚类算法 (Matched Filter, MF ； Adaptive Matched Filter, AMF ； color-magnitude filtered Voronoi Tessellation Technique, VTT) 在处理图像数据时的作用，发现 MF 在探测弱源时比较有效，而 AMF 在估计红移和密度时比较准确。将 MF 和 AMF 混合起来的模型称为 HMF 。 HMF 在背景均匀时优于 VTT ，在背景非均匀时也比 VTT 敏感；当对 SDSS 巡天的探测阈值选择适当时，两种算法的效果相当。 Louys 等人 [71] 对比了一系列的图像压缩方法：分形 (Fractal) 、小波 (wavelets) 、 PMT (Pyramidal Median Transform) 、JPEG 和一些软件包 HCOMPRESS 、FITSPRESS 、Mathematical Morphology ，发现没有一种方法是十全十美的。相比之下， PMT 对一般的天文图像具有较好的压缩能力；在压缩因子小于 40 的情况下， JPEG 是很好的方法。

4.7 天文学中数据挖掘技术所面临的挑战

天文学中，数据挖掘技术正面临着如下的挑战：

(1) 扩充数据挖掘算法：因为观测记录或观测次数的增长、每次观测参数的增长、分析观测数据的预测模型数增长，都对交互式反应和真实反应时间减少的要求加强，所以需要多种算法组合或开发新的算法。

(2) 应用于新的数据类型：如时间序列的数据、未组织的数据 (文本数据) 、半组织数据 (HTML 和 XML 文件) 、多媒体关联数据、多层次多度量单位的关联数据、集合数据。

(3) 开发新的分布式的数据挖掘算法：由于数据的分布特性和计算环境越来越普及，故必须开发与之匹配的新的数据挖掘系统和新的算法。

(4) 提高数据挖掘方法的容易度：包括提高数据挖掘自动化程度；提高用户界面以支持随机用户的浏览；提高大型分布数据的可视化程度；进一步开发用以管理数据挖掘的元数据的技术和系统；进一步开发恰当的语言和协议以支持随机提取数据；提高数据挖掘和知识发现的环境：从数据收集到数据加工、数据挖掘，再到可视化以及结果的评估和解释。

种种的挑战迫使数据挖掘技术不断地改进和提高，来支持单个数据挖掘者的研究、数据挖掘的基础学科的研究，支持多学科和交叉学科研究组研究重要的、基础的实用数据挖掘问题，提供大型的、分布式的数据挖掘恰当的实验场所。数据挖掘技术的发展仅靠天文学家是远远不够的，它需要来自计算机界、统计学界、数学界的科学家们的精诚合作；同时，巡天技术的迅猛发展、数据量的飞速增长，也需要新的数据存储方法、新的分析工具、强大的软硬件支持，以及更多的适应时代需要的科学家。

5 结 语

综上所述，数据挖掘和知识发现是一个利用各种分析工具在海量数据中发现模型和数据间关系的过程，并且可以通过这些模型和关系做出预测。数据挖掘的质量完全依赖于数据的数量、质量与算法的优劣。天文数据自身的复杂性特征是天文数据挖掘和知识发现理论不断发展和完善的内因，并在一定程度上左右着天文学理论前进的方向，因此在未来的一段时间内天文数据知识发现研究的主要任务仍然是解决天文数据中蕴藏的复杂性问题。而其它领域

(如模式识别、机器学习、统计理论) 出现的新理论、新方法，将是天文数据知识发现理论逐渐成熟的外在动力。面对种种挑战，数据挖掘技术会在效率和质量上得到充分的改善。未来的虚拟天文台将把分布的、海量的数据有机地结合起来，这就需要开发与之匹配的强而有效的数据挖掘工具，用以处理这些高精度的海量数据，有效地解决天文学中的“数据雪崩”，发现新类型的天体，并从结果中得出一些新的有意义的天体物理知识。

在数据挖掘和知识发现过程中，使用者 (天文学家) 的因素亦是至关重要的。可以说天文学家是联系天文数据与数据挖掘工具的枢纽。如何利用数据挖掘工具将天文数据转化为知识，是摆在每一位天文学家面前最实际的问题。作为天文学家既要懂得天文也要会使用数据挖掘工具，这不仅需要扎实的天文功底，而且还需要了解数学、统计学、计算机和模式识别等方面的知识。面对海量数据和各种挖掘技术，科学家的素养和实验路线的选取将直接影响数据分析的效率和新的发现。因此，每一位天文学家应积极努力地调整自己的知识结构，以适应新时代发展的需求。

参 考 文 献

1 Szalay A S, Brunner R J. In: Brian J McLean, Daniel A Golombek, Jeffrey J E Hayes *et al.* eds. Proc. of IAU colloq. 179, Dordrecht: Kluwer Academic Publishers, 1998: 455

2 Szalay A, Gray J. Science, 2001, 293: 203

3 http://www.us-vo.org/docs/nvo-proj.pdf

4 Fayyad U, Piatetsky-Shapiro G, Smyth P *et al.* eds. Advances in Knowledge Discovery and Data Mining, Boston: AAAI/MIT Press, 1996: 1

5 Babu G J, Feigelson E D. ASP Conf. Ser., 2001, 225: 272

6 Borne K D. In: Banday A J, Zaroubi S, Bartelmann M eds. Proc. of the MPA/ESO/MPE Workshop, Heidelberg: Springer-Verlag, 2001: 671

7 Thakar A R, Kunszt P Z, Szalay A S. ASP Conf. Ser., 2001, 225: 230

8 Moore R W. ASP Conf. Ser., 2001, 225: 257

9 Williams R. ASP Conf. Ser., 2001, 225: 302

10 Moore A W, Lee M S. Journal of Artificial Intelligence Research, 1998, 8: 67

11 Nichol R C, Miller C J, Connolly A *et al.* In: Banday A J, Zaroubi S, Bartelmann M eds. Proc. of the MPA/ESO/MPE Workshop, Heidelberg: Springer-Verlag, 2001: 613

12 Miller C J, Genovese C, Nichol R C *et al.* Ap. J., 2001, 122: 3492

13 Hopkins A M, Miller C J, Connolly A J *et al.* Ap. J., 2002, 123: 1086

14 Komarek P, Moore A. International Conference on Machine Learning, 2000, http://www.autonlab.org/pap.html

15 Anderson B, Moore A. Knowledge Discovery From Databases, 1998, http://www.autonlab.org/pap.html

16 Pelleg D, Moore A. International Conference on Machine Learning, 2000, http://www.autonlab. org/pap.html

17 Moore A. Proceeding of Advances in Neutral Information Processing Systems 11, 1999, http://www.autonlab.org/pap.html

18 Maino D, Farusi A, Baccigalupi C *et al.* M.N.R.A.S, 2001, preprint (astro-ph/0108362)

19 Bishop C M. Neural Networks for Pattern Recognition, Oxford UK: Oxford University Press, 1995

20 Kohonen T. Proc. IEEE, 1990, 78(9): 1464

21 Bazell D, Peng Y. Ap. J. Suppl., 1998, 116: 47

22 Andreon S, Gargiulo G, Longo G *et al.* M.N.R.A.S., 2000, 319: 700

23 Weir N, Fayyad U, Djorgovski S G *et al.* Publ. Astron. Soc. Pac., 1995, 107: 1243
24 Weir N, Fayyad U, Djorgovski S G. A. J., 1995, 109: 2401
25 Fayyad U, Smyth P, Weir N *et al.* J. Intel. Inf. Sys., 1995, 4: 7
26 Bertin E, Arnout S. Astron. Astrophys. Suppl. Ser., 1996, 117: 393
27 Storrie-Lombardi M C, Lahav O, Sodre L Jr *et al.* M.N.R.A.S., 1992, 259: 8
28 Lahav O, Naim A, Sodre L Jr *et al.* M.N.R.A.S., 1996, 283: 207
29 Bailer-Jones C A L, Irwin M, von Hippel T. M.N.R.A.S., 1998, 298: 361
30 Allende Prieto C, Rebolo R, Lopez R J G *et al.* A. J., 2000, 120: 1516
31 Weaver W B. Ap. J., 2000, 541: 298
32 Longo G, Tagliaferri R, Andreon S. In: Banday A J, Zaroubi S, Bartelmann M eds. Proc. of the MPA/ESO/MPE Workshop, Heidelberg: Springer-Verlag, 2001: 379
33 Babu G J, Feigelson E D. Astrostatistics, London: Chapman & Hall, 1996: 11
34 Babu G J, Feigelson E D. Statistical Challenges in Modern Astronomy II, New York: Springer-Verlag, 1997: 72
35 Murtagh F, Heck A. Multivariate Data Analysis, Dordrecht: Kluwer, 1987
36 Nichol R C, Connolly A J, Moore A W *et al.* ASP Conf. Ser., 2001, 225: 265
37 Connolly A J, Genovese C, Moore A W *et al.* 2000, preprint (astro-ph/0008187)
38 Adanti S, Battinelli P, Capuzzo-Dolcetta R *et al.* Astron. Astrophys. Suppl. Ser., 1994, 108: 395
39 Connolly A J, Szalay A S, Bershady M A *et al.* A. J., 1995, 110 (3): 1071
40 Connolly A J, Szalay A S. A. J., 1999, 117: 2052
41 Zhang Yanxia, Zhao Yongheng. Ap. J., 2002, 待发表
42 Burges C J C. Data Mining and Knowledge Discovery, 1998, 2: 121
43 Spiekermann G. Ap. J., 1992, 103: 2102
44 Mahonen P, Frantti T. Ap. J., 2000, 541: 261
45 Tenenbaum J B, de Silva V, Langford J C. Science, 2000, 290: 2319
46 Roweis S, Saul L K. Science, 2000, 290: 2323
47 Cortiglioni F, Mahonen P, Hakala P *et al.* Ap. J., 2001, 556: 937
48 Connolly A J, Castander F, Genovese C *et al.* In: Banday A J, Zaroubi S, Bartelmann M eds. Proc. of the MPA/ESO/MPE Workshop, Heidelberg: Springer-Verlag, 2001: 323
49 Naim A, Ratnatunga K U, Griffiths E. Ap. J. Suppl., 1997, 111: 357
50 Naim A, Lahav O, Sodre L Jr *et al.* M.N.R.A.S, 1995, 275: 567
51 Mahonen P, Hakala P J. Ap. J., 1995, 452: 77
52 Mahonen P, Frantti T. Ap. J., 2000, 541: 261
53 Murtagh F D. ASP Conf. Ser., 1992, 25: 265
54 Jarvis J F, Tyson J A. SPIE, 1979, 172: 422
55 Jarrett T H, Chester T, Cutri R *et al.* A. J., 2000, 119: 2498
56 Lahav O. In: Banday A J, Zaroubi S, Bartelmann M eds. Proc. of the MPA/ESO/MPE Workshop, Heidelberg: Springer-Verlag, 2001: 33
57 Baccigalupi C, Bedini L, Burigana C *et al.* M.N.R.A.S., 2000, 318: 769
58 Zaroubi S. M.N.R.A.S, 2002, 331: 901
59 Moretti A, Lazzati D, Campana S *et al.* Ap. J., 2002, 570: 502
60 Hoffman Y. In: Banday A J, Zaroubi S, Bartelmann M eds. Proc. of the MPA/ESO/MPE Workshop, Heidelberg: Springer-Verlag, 2001: 223
61 Chang Tzu-Ching, Refregier A. Ap. J., 2002, 570: 447
62 Odewahn S C, Cohen S H, Windhorst R A *et al.* Ap. J., 2002, 568: 539
63 Fruchter A S, Hook R N. Publ. Astron. Soc. Pac., 2002, 114: 144
64 Lasenby A, Barreiro B, Hobson M. In: Banday A J, Zaroubi S, Bartelmann M eds. Proc. of the MPA/ESO/MPE Workshop, Heidelberg: Springer-Verlag, 2001: 15

65 Hobson M P, Jones A W, Lasenby A N *et al.* M.N.R.A.S., 1998, 300: 1
66 Sanz J L, Barreiro R B, Cayon L *et al.* Astron. Astrophys., 1999, 140: 99
67 Miller C J, Nichol R C. A. J., 2001, 555: 38
68 Szapudi I, Quinn T, Stadel J *et al.* Ap. J., 1999, 517: 54
69 Starck J L, Murtagh F, Querre P *et al.* Astron. Astrophys., 2001, 368: 730
70 Kim R S J, Kepner J V, Postman M *et al.* Ap. J., 2002, 123: 20
71 Louys M, Starck J L, Bonnarel F *et al.* Astron. Astrophys. Suppl. Ser., 1999, 136: 579

Data Mining and Knowledge Discovery in Database of Astronomy

Zhang Yanxia Zhao Yongheng Cui Chenzhou
(*National Astronomical Observatories, Chinese Academy of Sciences, Beijing 100012*)

Abstract

The necessity of data mining and knowledge discovery in database of astronomy, as well as its development in recent years, process and tasks, are reviewed. The complex characteristics of present astronomical data are analyzed, and the science requirements of data mining in astronomy are introduced. Simultaneously, the research development, hot topic and challenges in the filed of data mining and knowledge discovery in database are summarized. Its existence and development will have a prospect of wide applications in the 21st century, a great push to astronomy and also provide new challenges of knowledge and technology to astronomers. In addition, the successful applications of data mining technology in the Virtual Observatory is an important factor for its sufficient exertion.

Key words data mining—knowledge discovery in database—methods—data analysis