文章编号: 1000-8349(2010)02-112-16

述评

聚类算法在天文学中的应用

严太生 [1,2]，张彦霞 [1]，赵永恒 [1]，李 冀 [2]

(1. 中国科学院光学天文重点实验室 国家天文台，北京 100012; 2. 河北师范大学 物理学院，石家庄 050016)

摘要： 聚类算法是数据挖掘中用来发现数据分布和隐含模式的一种重要算法，它把大量数据点的集合分成若干类，使得每个类中的数据最大程度地相似，而不同类中的数据最大程度地不同。尤其对于大样本，在多参量和类别未知的情况下，该方法更为简洁有效。为了更好地使用这些算法，对数据挖掘领域的聚类分析方法及代表算法进行了讨论，阐述了数据挖掘对聚类算法的典型要求，并基于这些要求对数据挖掘中常用的聚类算法作了概括，以便于人们更容易、更快速地选择一种适用于具体问题的聚类算法。综述了数据挖掘中聚类算法的分类和原理以及常用的聚类算法在天文学中的具体应用，分析了它们各自的性能，并指出了其今后的发展趋势。

关 键 词：天文学；聚类算法；数据挖掘

中图分类号：P152 **文献标识码**：A

1 引 言

随着大型望远镜的精度和深度不断提高，特别是巡天望远镜的发展，天文数据急剧增加，目前的数据总量已经达到 10^{15} 量级，为探索各类天体和天文现象的物理本质提供了强有力的支撑。面对天文学“数据雪崩”和“信息爆炸”时代的到来，为了对这些数据进行科学有效地处理，传统的人工技术已经远远不够，这样就需要一套工具能对海量的数据自动地进行预处理、特征选择、数据挖掘和对结果解释评估等一系列工作。在此背景下，如何科学有效地利用这些来自大型数字巡天和数据库的海量数据，如何迅速准确地从这些海量数据中有效地提取所需要的知识，如何从数以亿计的天体或天文数据中进行科学发现，这是摆在天文学家面前不可回避的问题，直接影响着天文学的发展和研究进程。由于天文数据本身具有非线性、高维性和类别不确定性等特点，一般的数据分析方法不能解决，这就需要一些新的分析工具来

收稿日期：2009-05-13； **修回日期**：2009-11-24

基金项目：国家自然科学基金 (10778724), 863 资助项目 (2006AA01A120)，中国科学院国家天文台青年人才基金项目资助

解决，引入适合天文发展需要的非监督性质的算法，充分有效地从数据中挖掘出天文学家感兴趣和有意义的天体和天文现象，将有助于推动天文学理论的更进一步发展和完善。于是聚类算法应运而生，并被广泛应用于天文学的各个方面。

2 聚类算法及相关概念

聚类的定义最早是由 Everitt [1] 在 1974 年给出的，它是将物理或抽象对象的集合分组成为有类似的对象组成的多个类的过程。它要划分的类是未知的。由聚类所生成的簇是一组数据对象的集合，这些对象与同一个簇中的对象彼此相似，与其他簇中的对象相异。

为了在天文学中能很好地应用聚类算法，必须知道天文学对其要求及其聚类的过程。

2.1 天文学应用对聚类算法的要求

由于天文学中的数据非常复杂，具有海量性、高维性、非线性等特点，所以应用于天文学中的聚类算法要求比较苛刻，下面给出几条主要的要求。

1) 海量性：天文数据量庞大，所以聚类算法应能够处理海量数据；

2) 高维性：天文数据特别是光谱数据一般都有几十维上百维，因此聚类算法应具有处理高维数据的能力；

3) 非线性：聚类算法应具有处理天文学中非线性数据的能力；

4) 噪声缺值性：天文学中的原始数据一般都受噪声、宇宙线等污染，且由于仪器设备等的限制，有很多的缺值数据，所以聚类算法要能够处理缺值数据和脏数据 (指被宇宙线等污染的天体光谱数据) ；

5) 不敏感性：要求聚类算法对天文数据输入顺序不敏感；

6) 可解释性和可用性：聚类算法对天文数据处理的结果是可解释的、可理解的和可用的。只有这样，聚类算法才能更好地被天文学家理解和接受，从而可以用其分析结果去预测其它的数据。

2.2 聚类过程

聚类过程粗略地分为三步 [2] ：数据准备、数据挖掘，以及结果的解释与评估。其过程如图 1 所示。

(1) 数据准备

数据准备就是对被聚类的数据进行定义、预处理和表示，使它适合于特定的聚类方法。它是聚类过程中的第一个重要步骤，在整个聚类中起举足轻重的作用，直接关系到聚类结果的质量。它主要包括如下三个过程：

①数据的选择

搜索所有与聚类对象有关的内部和外部数据信息，并从中选择一个数据集或在多个数据集子集上聚焦，挑选适用于聚类的数据。

②数据的预处理

去除噪声数据和无关数据，去除缺值数据域，考虑时间顺序和数据变换等。提高数据的质量，为进一步的分析做准备，并确定将要进行的聚类操作类型。

③数据的转换

找到数据的特征表示，用维变换或转换的方法，减小变量的数目或找到数据的不变式，将数据转换成一个针对聚类算法建立的分析模型，而建立一个真正适合聚类算法的分析模型是聚类的关键。

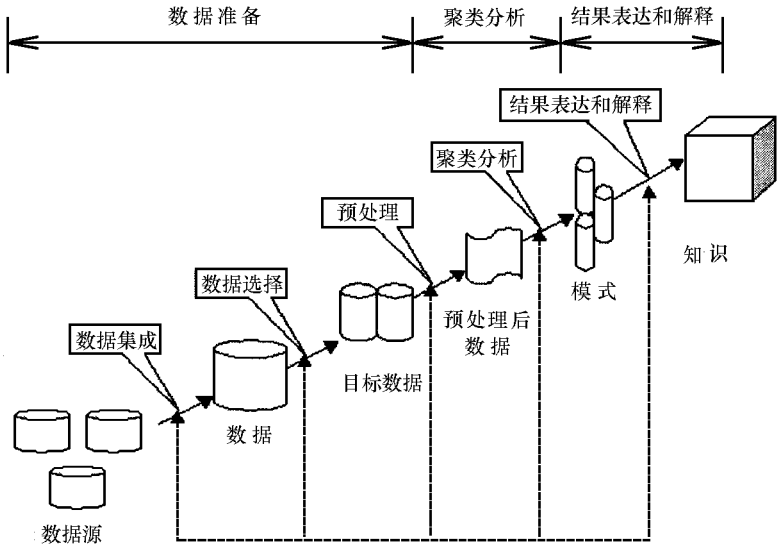

图 1 聚类分析流程图 [2]

(2) 数据挖掘 (聚类)

对预处理结果和变换后的数据进行聚类。选择某个或几个特定的聚类算法用于搜索数据中的模式。这些工作都是由算法去自动完成的。然后分析搜索结果产生一个特定的感兴趣的模式和数据集。

(3) 结果的解释与评估

解释并评估聚类结果的正确率。解释某个发现的模式，去除多余的无关的模式，转化为有用的模式，以使结果可理解和解释。其使用的方法一般由聚类操纵者决定，通常使用可视化和知识表示技术。

3 聚类算法的分类及原理

没有任何一种聚类算法可以普遍适用于揭示各种多维数据集所呈现出来的多种多样的结构 [3] 。聚类算法的选择取决于数据的类型、聚类的目的和应用。如果聚类算法被用作描述或探查的工具，可以对同样的数据尝试多种算法，以发现数据可能揭示的结果。大体上，主要的聚类技术可以划分为如下几类。

3.1 划分聚类算法

划分聚类算法 [4] 需要预先指定聚类数目或聚类中心，通过反复迭代运算，逐步降低目标函数的误差值，当目标函数值收敛时，得到最终的聚类结果。它可以应用于入侵检测 [5] 和孤立点聚类 [6] 等。

划分聚类算法的思想：给定一个包含 n 个对象的元组的数据库，一个划分方法构建数据的 k 个划分，每个划分表示一个聚簇，并且 $k \leq n$ 。即它将数据划分为 k 个组，同时满足如下要求： (1) 每个组至少包含一个对象； (2) 每个对象必须属于且只属于一个组。实际中，绝大多数采用了下面两个启发式方法： (1) k 均值算法，在该算法中，每个簇用该簇中对象的平均值来表示； (2) k 中心算法，在该算法中，每个簇用接近聚类中心的一个对象来表示。

典型的划分聚类算法有 k-means [7] 、 k-medoids [8] 和 CLARANS [6] 。 k-means (k 均值) 算法是一种基于样本间相似性度量的间接聚类方法，相似度的计算根据一个簇中对象的平均值 (被看作簇的重心) 来进行。k-medoids (k 中心点，基于有代表性的对象的技术) 算法选用簇中位置最中心的对象，即中心点 (medoid) 作为参照。 CLARANS(大型数据库中的划分方法) 算法是划分方法中基于随机搜索的大型应用聚类算法，在搜索的每一步都带一定随机性的选取一个样本。它拓展了数据处理量的伸缩范围，具有较好的聚类效果，但计算效率较低，且对数据输入顺序敏感，只能聚类凸状或球型边界。

3.2 层次聚类算法

层次式聚类算法 [9] 将给定的数据对象组成一棵聚类的树，对它进行层次的分解。根据层次分解是自下向上的还是自上向下形成的，层次的聚类方法可以进一步分为凝聚的 (agglomerative) 层次聚类算法和分裂的 (divisive) 层次聚类算法两种。它们可以应用于人脸识别 [10] 等。

凝聚的层次聚类思想：首先将每个对象作为单独的一个簇，然后合并这些原子簇为越来越大的簇，直到所有的对像都在一个簇中，或者达到某个终止条件。

分裂的层次聚类思想：首先将所有的对象置于一个簇中。然后逐渐细分为越来越小的簇，直到每个对象在单独的一个簇中，或者达到一个终止条件。

典型的层次聚类算法有 BIRCH [11] 、 CURE [12] 、 ROCK [13] 和 Chameleon [14] 。 BIRCH(利用层次方法的平衡迭代归约和聚类) 算法是一种较为灵活的递增式 (incremental) 聚类方法，能根据内存的配置大小而自动调整程序对内存的需要，它有良好的算法伸缩性 (scalability) 、数据输入顺序不敏感性、较好的聚类效果等优点。 CURE(利用代表点聚类) 算法是一种针对大型数据库的高效的聚类算法，它采用了用多个点代表一个簇的方法，可以较好地处理任意形状数据和异常数据。 ROCK(实用于分类属性聚类) 算法根据相似度阈值和共享近邻的概念从给定的数据相似度矩阵构建一个稀疏的图，然后在这个稀疏图上执行一个层次聚类算法。 Chameleon(利用动态模型的层次聚类) 算法是一个在层次聚类中采用动态模型的层次聚类算法。

3.3 基于密度聚类算法

绝大多数划分方法是对对象之间的距离进行聚类。这样的方法只能发现球状的簇，而在发现任意形状的簇上遇到了困难。为了解决这个问题，提出了基于密度聚类算法 [15] ，它是通过数据密度 (单位区域内的实例数) 来发现任意形状的类簇，应用于调制识别 [16] 和图像分割 [17] 等。

基于密度聚类算法的思想：只要邻近区域的密度 (对象或数据点的数目) 超出了某个阀值，就继续聚类。也就是说，对给定类中的每个数据点，在一个给定范围的区域中必须至少含有一定数目的点。这样的方法可以用来过滤“噪声”孤立点数据，发现任意形状的簇。

典型的基于密度聚类算法有 DBSCAN [18] 、OPTICS [19] 和 DENCLUE [20] 。DBSCAN(基于高密度连接区域的密度聚类) 算法通过不断生长足够高密度区域来进行聚类; 它能从含有噪声的空间数据库中发现任意形状的聚类。此方法将一个聚类定义为一组 “密度连接” 的点集。OPTICS(通过对象排序识别聚类的结构) 算法并不明确产生一个聚类, 而是为自动交互的聚类分析计算出一个增强聚类顺序。 DENCLUE(基于密度分布函数的聚类) 算法通过表征对邻居影响的 “影响函数” 的总和来表示整个数据空间的密度, 类是通过确定某个区域密度的最高峰来确定。

3.4 基于网格聚类算法

基于网格聚类算法 [21] 采用一个多分辨率数据结构, 围绕模式组织由矩形块划分的值空间, 基于块的分布信息实现模式聚类。这种方法的主要优点是处理速度快, 其处理时间独立与数据对象的数目, 只与量化空间中的每一维的单元数目有关。主要应用于案例检索 [22] 等。

基于网格聚类算法思想: 把对象空间量化为有限数目的单元, 形成一个网格结构。所有的聚类操作都在这个网格结构 (即量化的空间) 上进行。

典型的基于网格聚类算法有 STING [23] 、WaveCluster [24] 和 CLIQUE [25] 。STING(统计信息网格) 算法由上而下将数据空间切割成格子状, 并通过树状结构呈现出来, 然后利用广度搜索将格子内的群聚类。 WaveCluster(采用小波变换聚类) 算法是一种多分辨率的聚类算法, 它通过在数据空间上强加一个网格结构来汇总数据, 然后采用一种小波变换来变换原特征空间, 在变换后的空间进行聚类。这种聚类算法速度很快, 可以实现并行化。 CLIQUE(聚类高维空间) 算法综合了基于密度和基于网格的聚类算法, 区分空间中稀疏的和密集的区域, 以发现数据集合的全局分布模式进行聚类。

3.5 基于模型聚类算法

基于模型聚类算法试图优化给定的数据和某些数学模型之间的适应性, 为每个簇假定了一个模型, 寻找数据对给定的模型的最佳拟合。这样的方法经常基于这样的假设: 数据是根据潜在的概率分布生成的。它可以应用于文本聚类 [26] 等。

基于模型聚类算法思想: 通过构建反映数据点空间分布密度函数来定位聚类。它也基于标准的统计数字自动决定聚类的数目, 考虑 “噪声” 数据或孤立点, 从而产生健壮的聚类方法。

典型的基于模型聚类算法有 COBWEB [27] 和 SOM [28] 。 COBWEB(概念聚类) 算法是以一个分类树的形式创建层次聚类, 它的输入对象用分类属性 — 值对来描述。 SOM(自组织映射) 算法是将多维数据映射到低维规则网格中, 可以有效地进行大规模聚类。

3.6 基于约束聚类算法

现实中的聚类问题往往是具备多种约束条件的, 然而, 由于在处理过程中不能准确表达相应的约束条件, 不能很好地利用约束知识进行推理以及不能有效利用动态的约束条件, 使得这一方法无法得到广泛的推广和应用。这里的约束可以是对个体对象的约束, 也可以是对聚类参数的约束, 它们均来自相关领域的经验知识。该方法的一个重要应用是对存在障碍数据的二维空间数据进行聚类。

基于约束聚类算法思想: 用两点之间的障碍距离取代了一般的欧氏距离来计算其间的最

小距离。

典型的基于约束聚类算法有 COD [29](障碍距离聚类) 算法，它是对障碍数据的二维空间数据进行聚类。

为了能更清晰地了解各种聚类算法，以便于人们更容易、更快速地选择合适的方法，表 1 列举了聚类算法的基本类型及其代表算法，前面六种就是文章前面详细介绍的传统聚类算法，而后面五种是近年来发展的新型的聚类算法。

表 1　聚类算法的基本类型及其代表算法

基本类型			代表算法
传统聚类算法	层次方法	聚合聚类	CURE 、 ROCK 、 Chameleon
		分裂聚类	BIRCH
	划分方法		k-means 、 k-medoids 、 CLARANS
	基于密度的方法		DBSCAN 、 OPTICS 、 DENCLUE
	基于网格的方法		STING 、 WaveCluster 、 CLIQUE
	基于模型的方法	统计学方法	COBWEB
		神经网络方法	SOM
	基于约束的方法		COD
新聚类算法	基于模糊的聚类方法		FCM
	基于粒子的聚类方法		(不成熟)
	量子聚类		QC
	核聚类		FC
	谱聚类	迭代谱聚类	PM 、 SM 、 SLH 、 Mcut
		多路谱聚类	NJW 、 MS

4　聚类算法在天文学中的应用

聚类算法是数据挖掘、模式识别等研究方向的重要研究内容之一，在识别数据的内在结构方面具有极其重要的作用。聚类算法主要应用于模式识别中的语音识别 [30] 、字符识别等，机器学习中的聚类算法应用于图像分割 [31] 和机器视觉，图像处理中聚类算法用于图像压缩 [32] 和信息检索 [33] 。此外，聚类分析在天文学、生物学、考古学、地质学以及地理学中都有重要应用。本文只介绍它在天文学领域中的具体应用。

4.1　恒星 / 星系分类

恒星 / 星系的分类是天文学的基本任务之一，在假设星系的图像比恒星的更延展或者更模糊前提下，根据恒星和星系在不同波段的表现性质的不同，应用不同的方法将它们各自区分开来。这对人们了解恒星和星系形成与演化的历史以及发现特殊天体都具有重要的研究价值。尤其对现在日益发展的大型巡天计划及由此产生的海量数据而言，如何将天体自动分类显得尤为重要。

对恒星 / 星系分类可以采取两种截然不同的方法：第一种是根据点源 (恒星) 和展源 (星

系) 的不同表现来研究 r 波段的点扩散函数 (point spread function，简称 PSF) 星等与模型星等的差值分布情况，给出它们的划分标准。 Strauss 等人 [34] 选择 $r^*_{\rm PSF}-r^*_{\rm model}$=0.3(如图 2 所示， * 代表红化矫正) 作为分类方法对 13 772 个 r 波段的 Petrosian 星等 $r^*_{\rm P}<17.8$ mag 的天体进行分类，结果表明没有明显的误分，正确率达到 98% 以上。虽然这种截断方法得到了很高的正确率，但它只能实用于这样低维的有非常明显聚类特征的数据中，并不能充分利用所给的信息，带有很大的偶然性。

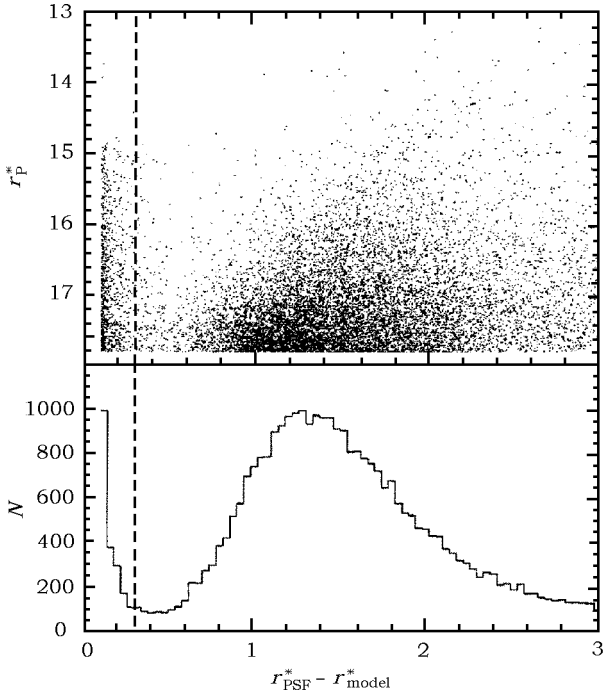

图 2 Strauss 恒星 / 星系分类 $r^*_{\rm PSF}-r^*_{\rm model}$ 星等图和直方图 [34]

为了弥补这种方法的不足，提出了应用算法对天体进行自动聚类 / 分类，这就是第二种方法。目前，已有许多研究者在这方面进行了研究与探索工作。 Mähönen 等人 [35] 应用 FCM 算法 (模糊 c 均值聚类，用隶属度确定每个数据点属于某个聚类的程度的一种聚类算法。首先随机选择 c 个初始聚类中心，然后根据最小距离原则将每个样本分配到某一类中，之后不断迭代计算各类的聚类中心并依据新的聚类中心调整聚类情况，直到迭代收敛) 把 POSS 产生的 9 245 个天体自动聚成两类，得到合理的结果，表 2 给出了聚类的详细结果。同时，与神经网络的 BP

表 2 Mähönen 应用三种算法聚类结果 [35]

Size	Fuzzy	BP	SOM
200	77	84	86
400	84	92	93
600	96	96	97
800	94	98	99
1000	95	98	99
1200	99	99	99

算法 (误差反向传播算法， BP 算法由信息的正向传递与误差的反向传播两部分组成。在正向传递过程中，输入信息从输入层经隐含层逐层计算传向输出层，每一层神经元的状态只影响下一层神经元的状态。如果在输出层没有得到期望的输出，则计算输出层的误差变化值，然后转向反向传播，通过网络将误差信号沿原来的连接通路反传回来，修改各层神经元的权值直至达到期望目标) 和 SOM 算法比较，发现神经网络算法在少量数据和大星等天体中分类效果更优。在大量数据中，它们的正确率相当，都高达 99% ，但模糊聚类算法能直接估计出分类的可靠性，从而得到比较全面的分类信息。通过测试发现使用少量参数就能够得出较好的分类结果，但使用的参数越多，得到的可靠性越高。自动聚类 / 分类算法对数据本身没有要求，它能充分利用所给的信息，对大量的、复杂非线性的、高维的数据进行分类，其结果具有物理意义，有很好的可解释性和可用性，因此能很好地满足天文学对算法的要求。

4.2　星系形态分类

星系是宇宙的基本组成单元，而形态类型是它的一个基本特性。星系形态分类是天文学家根据外观将星系系统地分成不同的类别，是更好地了解星系物理性质的首要步骤。自从 Hubble [36]1926 年提出星系形态分类，它就一直是天文学中一个重要的课题。以前通常是观察星系的图像来进行分类的，但这需要有一定的经验和专业知识，同时也是费时费力的。随着大型望远镜和大规模的巡天项目获得大量的星系数据，对星系形态进行自动分类变得越来越重要。

Strateva 等人 [37] 分析了 SDSS 的星系数据分布规律，发现 u^*-r^* 颜色有很强的双峰性，如图 3 所示。选择 u^*-r^* =2.22 作为早型星系与晚型星系分类的标准。并应用 AutoClass 算法 (自动聚类算法，它是一种贝叶斯分类方法，通过对数据进行处理，计算出每条数据属于每个类别的几率值) 将数据进行分类。对 25 000 个星系的颜色进行聚类，结果显示分类的 82% 与前面的标准一致， 5% 离群，剩下的与 82% 部分重叠。 AutoClass 分类结果就很好验证了 u^*-r^* 的双峰性用作两类星系分类标准的有效性和可行性。 Ball 等人 [38] 给出了星系的包括

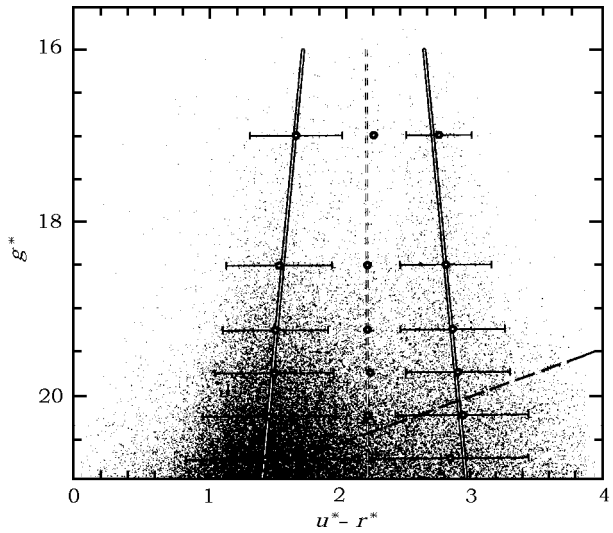

图 3　Strateva 星系 u^*-r^* 的颜色 – 星等图呈现双峰特性 [37]

汇聚指数 c=R50/R90 、 sersic 指数 n 、表面亮度、颜色、光谱类型 eClass 等的具体分布情况，为运用算法对多维空间分类的参数选择提供了科学依据。

星系的形态分类可以有助于加深人们对星系的结构及其演化过程理解，在研究宇宙大尺度方面也起到很重要的作用。星系的聚类可以有效地追踪宇宙的质量分布规律。通过测量不同星系聚类的质量， Martínez 等人 [39] 研究表明，星系遵循着质量分布的演化规则，而 Hartigan 等人 [40] 更是表明，发现星系聚类等同于发现等密度线聚类。所有这些都说明星系的聚类对于研究星系速度弥散 [41] 、宇宙学参数 [42,43] 等物理参数分布规律相关方面都有及其重要的指导意义。

4.3 探索 SDSS 恒星 (变星) 数据库

恒星是星系的基本组成单元，其结构与演化理论是天体物理的理论基础。而变星对人们研究恒星的结构、演化等基本物理过程，对测量天体的距离、年龄等基本物理量，对联系天体物理中的三个层次 (恒星、星系、宇宙) 起着极其重要的作用，但是人们对于变星的理解还非常的肤浅 [44] 。

随着产生多历元测光的巡天等设备问世，人们获得了大量的变星数据，这对变星的进一步理解提供了丰富的资料，而对这些数据的分析要通过算法自动进行。 Eyer 等人 [45] 应用 AutoClass 聚类算法对 ASAS (All-Sky Automated Survey ，全天自动巡天)1-2 的 1 700 多个变星自动聚成六类：食变双星、天琴座 RR 型变星、造父变星、微幅红变星、半规则变星和 Mira 变星，其中有 302 个周期性变星和 1 429 个半周期性变星，聚类结果见图 4 。结果发现 83% 的变星落在红巨星范围内，即图 4(b) 。从图 4(b) 可以看出，大部分红巨星光谱型从左低段的 K 型到右高段的 M 型。在大幅度长周期部分，发现了著名的 Mira 变星。由于星等的限制，天琴座 RR 型变星因为太暗而难以观测，因此这类天体非常稀少。分类的误差在 7% 左右，表明这种方法能有效地处理变星巡天数据，从而为更详细的变星研究提供了一种强有力的武器。

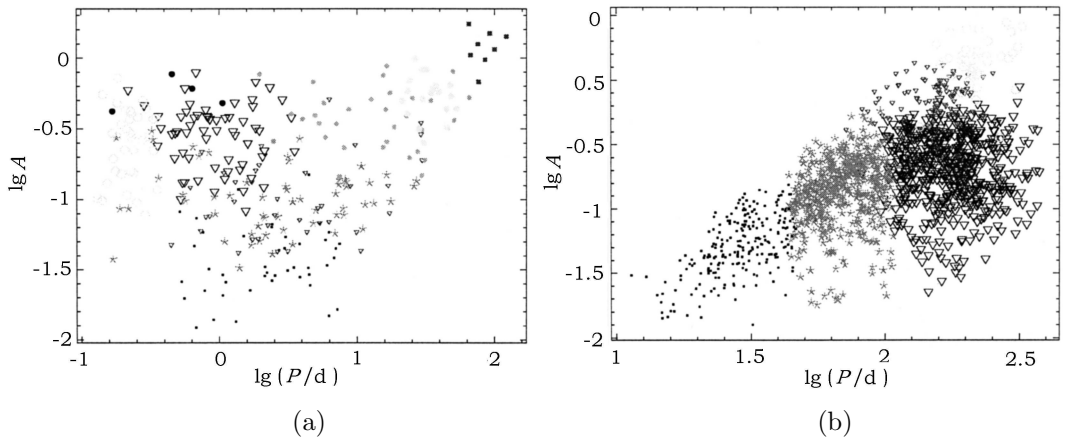

图 4 Eyer 的变星聚类的周期 – 幅度对数关系图 [45]

(a) 不包括红巨星的聚类分布 (大空心三角形、空心圆环和小空心三角形代表食变双星，大黑点代表天琴座 RR 型变星，小黑点和圆环代表造父变星，大实心正方形代表半规则变星，小实心正方形代表微幅变星)； (b) 红巨星的聚类分布结果 (小实心正方形代表微幅红巨星，空心三角形代表半规则变星，圆环代表 Mira 变星) 。

4.4　光谱分类

光谱分类也是天文学中一个重要的研究课题，通常是将未知的光谱型和标准的光谱作比较，从而识别它的类别，但这种方法效率非常低，只能处理小量的数据，同时还受人的主观因素影响，所以很难形成统一的模式。随着大型望远镜开发，巡天的广度和深度都有很大的提高，观测数据爆炸式增长，这种人工的方法已经远远不能满足目前的要求了，所以光谱的自动分类和聚类变得非常重要。

和测光数据相比，光谱数据所给的信息更全面、准确，信息量更大，同时数据维数高，所以处理起来比测光数据困难得多。Hojnacki 等人 [46] 使用 PCA(主成分分析，从混合信号中求出主成分，次成分与主成分相对，它是混合信号中能量最小的成分，被认为是不重要的或是噪声有关的信号，把确定次成分的方法称为次成分分析 MCA) 和 k-means 等聚类算法把来源于 COUP (Chandra Orion Ultradeep Project observation) 的 444 个高质量 CCD 图像 X 射线光谱聚成 17 类，并对各个类别进行分析，获得很好的结果，如图 5 所示。然后用 1 333 个 NGC 的 X 射线源进行检测，得到正确率为 94% ，说明这种方法可以胜任此类光谱精确的分类。

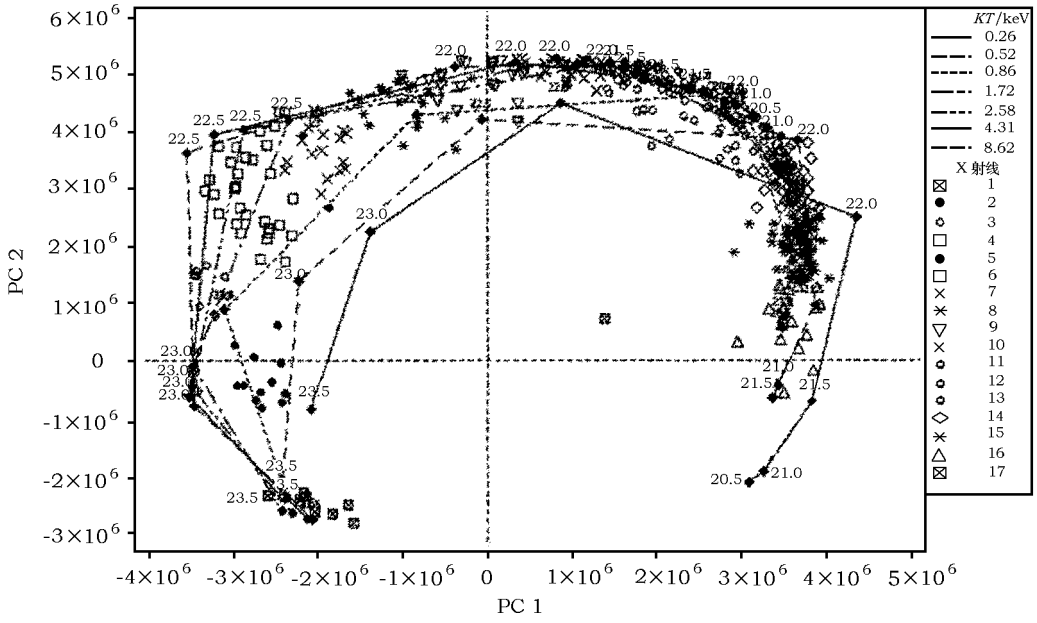

图 5　COUP 的 444 个 X 射线光谱聚类成 17 类并通过 PCA 得到 PC1-PC2 的分布图 [46]

4.5　时间序列数据分析

随着观测设备的数量、灵敏度等逐渐提高，发现有大量的变体存在于宇宙中并越来越多，它们包括可以预知行为的恒星 (如造父变星) 、不可预知行为的天体 (如活动星系核) ，以及可以预测的但不规则变化模式的天体 (如 X 射线双星等) 。时间序列数据分析最相关的问题是在不均衡的抽样中有多少时间序列被收集以及怎样有效地发现时间序列模式，而不是研究时间序列曲线的可能周期数这个自然结果。这些时间序列数据对于研究变星、活动星系核等类型的天体特别重要。时间序列信号分析是一个从地质学到天文学等很多领域都非常相关的课题 [47,48] 。

天文学中常用的傅里叶变换和功率谱法可以很好地用于发现固定周期时间序列模式 [49]。然而，它们对于不规则时间序列的分析却无能为力，也不能作为预测工具对这些数据进行优化。为了克服这些缺点， Perlman 等人 [50] 应用能够发现和优化不规则变化模式的聚类算法来分析时间序列。首先通过一个滑动窗口来收集时间序列数据。然后运用贪心聚类算法 (greedy clustering algorithm ，简称 GCA ，可直接确定彼此相关性低，且重要程度依次降低的聚类中心) 对这些数据进行聚类。一旦获得一个适宜的结果，这些聚类就被当作时间序列的基本形状，整个时间序列由这些基本形状重叠和 / 或结合组成。接下来也是最重要的是寻找两个时间序列聚类之间有趣的、随机的规则。图 6 显示的是用聚类算法对时间序列分析预测的流程图。研究结果表明时间序列的长期行为是由短期的变更叠生的，但是仅仅发生在亮的天体上 [51] ，它对预期短期的规则特别有效。长期的规则既可以通过改变窗口尺度来发现，也可以通过平滑法来实现。这种方法简单可靠，对频率间隔不敏感。

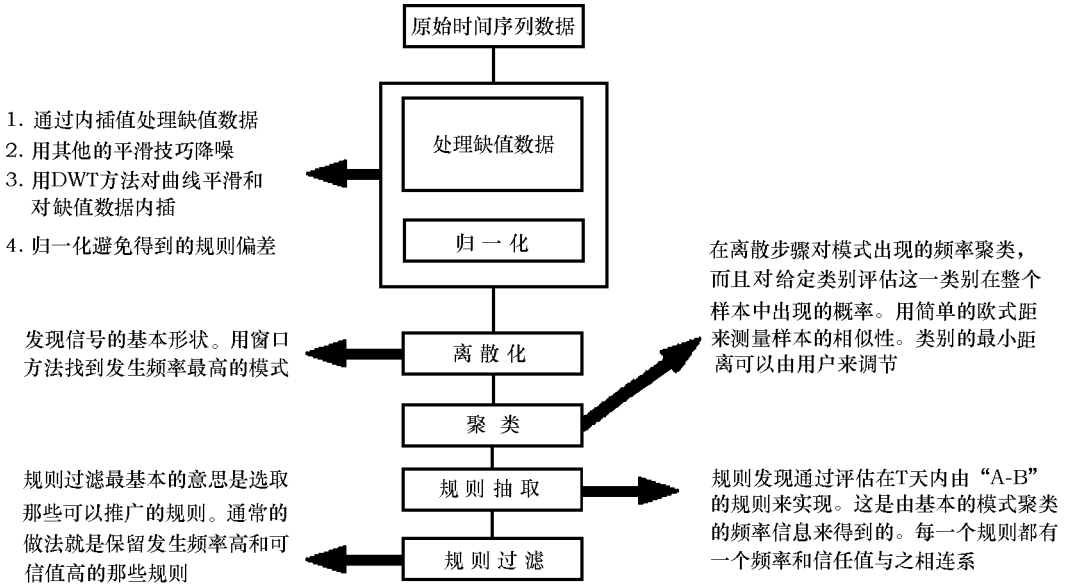

图 6 Perlman 的聚类算法对时间序列分析预测的流程图 [51]

4.6 缺值数据的恢复

由于观测设备本身缺陷和观测环境的影响，天文学等领域观测数据中存在大量的缺值数据。如何对这些数据进行插补或恢复是数据挖掘中面临的一个客观的且具有挑战性的问题，它可能会影响聚类的质量和性能。然而，现有的大部分算法都是在假设没有缺值的情况下设计的，因此它们不能很好地应用于有缺值的领域。为了能对缺值数据进行有效的处理， Green 等人 [52] 制定了两套处理缺值数据的可行方案： (1) 数据插补法，估计一个数值 (平均值、最大值或最小值等) 填补到缺值数据中； (2) 数据忽略法，忽略缺值数据。这些都是在假定缺值数据是随机产生的前提下进行的，表 3 是 Wagstaff 等人 [53] 对一些常用数据忽略法和数据插补法进行的比较。尽管这些方法在一些领域适用，但是它们不能很好地适用于天文数据，因为在天文学中缺值很可能有它的物理意义 [54] 。

为了克服上述方法的缺点， Wagstaff 等人 [53] 应用软约束 k-means 算法 (K-means Clustering with Soft Constraints，简称 KSC，结合归一化方差和约束异常值选择一个修改的目标函数) 对缺值数据进行恢复，避免了插值等人为因素产生的影响。在这里，将数据特性分成两类。对全部观测特性数据进行聚类，而用部分的观测特性数据 (包括缺值数据) 产生一系列聚类算法的约束，将部分观测特性数据的信息结合起来作为全部观测特性数据的信息增补。与插补法比较， KSC 能将全部观测特性数据聚类分配的影响减到最小。当可提供特性的观测数据非常少时，能推测缺值数据的信息就非常少，这时插补法很难可靠。表 3 是一些常用方法的比较，可以间接地发现 KSC 算法在效率、可靠性和稳定性等方面更优。

表 3　Wagstaff 的一些常用数据忽略法和数据插补法的比较 [53]

方　法	优　点	缺　点
特征忽略：不考虑缺值的特征	简单	研究对象的信息损失
对象忽略：不考虑具有缺值的对象	简单	研究对象的损失
平均值替代：用数据集的平均值代替每一个缺值数据	简单	可能不正确；一般平均值从来不会出现
几率值替代：按照数据的分布，随机取值替代缺值数据	推导的值是“真实的” (实际观测值)	推导的值可以与研究对象无关联
近邻值替代：用邻近对象的属性值替代缺值数据	推导的值可能是最好的猜测值	推导的值仍然可能不准确 (即非观测值)

4.7　其他应用

聚类算法在天文学中的应用非常广泛，除了上面介绍的应用外，还在很多领域有重要的应用。 Xu 等人 [55] 研究表明，如果满足一致性， k-means 算法可以将任意的形状聚类得非常紧凑和简洁。与谱聚类特性比较，它不需要解决本征值问题，而且效率更高。 Hakkila 等人 [56] 应用四种聚类算法 (ESX 算法、 EM 算法、 k-means 算法和 kohonen 神经网络算法) 将 BATSE γ 射线暴分别聚类成两类、三类和四类，结果表明在短且微弱射线和长且强烈射线分界地方的数据偏离中心且没有明显特性。 Fuentes 等人 [57] 选择三种不同的光谱特性 (谱指数、谱线和光谱) 数据并应用 distance-weighted 3-nearest-neighbor 算法预测恒星大气参数 Teff, lg g 和 [Fe/H]，实验结果表明选择谱指数和谱线的准确率非常相似，都优于选择光谱得到的准确率。 Baccigalupi 等人 [58] 使用 ICA 算法模拟 4 个频率上的微波背景辐射 (CMB) 、尘埃热辐射、星系同步加速和核外射电源，模拟的结果见表 4 。这些结果都说明 ICA 算法适应于各种天体物理应用。

表 4 Baccigalupi 等使用 ICA 算法模拟四种源的模拟结果 [58]

频率	射电源		CMB		同步加速辐射		尘埃热辐射	
/GHz	输入	输出	输入	输出	输入	输出	输入	输出
100	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
70	1.97	1.95	1.14	1.14	2.81	1.36	0.68	0.93
44	4.76	4.70	1.22	1.23	10.8	1.72	0.35	1.93
30	9.86	9.70	1.26	1.26	32.8	−12.0	0.19	3.77

5 聚类算法比较

为了能够选择合适的聚类算法，本文对目前的一些聚类算法进行了比较。而聚类算法的比较研究是基于下面 5 个标准。

(1) 是否适用于大数据量，算法的效率是否满足大数据量高复杂性的要求;

(2) 是否能应付不同的数据类型，能否处理符号属性;

(3) 是否能发现不同类型的聚类;

(4) 是否能应付脏数据或异常数据;

(5) 是否对数据的输入顺序不敏感。

比较结果见表 5 所示。从表 4 可以看出 BRICH 、 CURE 和 SOFM 三种算法的效率都比较高，对异常数据不敏感，而且对数据的输入顺序也不太敏感，其中以 CURE 实用范围最广，效率最高。

表 5 聚类算法比较结果表

算法	算法效率	适合的数据类型	发现的聚类类型	对脏的或异常数据的敏感性	对数据输入顺序的敏感性
BRICH	高	数值	凸形或球形	不敏感	不太敏感
DBSCAN	一般	数值	任意形状	敏感	敏感
CURE	高	数值	任意形状	不敏感	不太敏感
K-pototype	一般	数值或符号	凸形或球形	敏感	一般
CLARANS	较低	数值	凸形或球形	不敏感	非常敏感
CLIQUE	较低	数值	凸形或球形	一般	不敏感
k-means	较高	数值	凸形或球形	敏感	敏感
SOFM	较高	数值	任意形状	不敏感	不敏感
PAM	较低	数值	任意形状	不敏感	不敏感

6 总结与展望

聚类算法在天文学研究中具有广泛的应用前景, 今后的发展也将面临着越来越多的挑战。

由于天文数据本身具有高维性、复杂性、动态性以及容易达到大规模的特性，对聚类算法的使用还应该更多地考虑以下几个方面的内容：

(1) 融合不同的聚类思想形成新的聚类算法，从而综合利用不同聚类算法的优点。

(2) 处理大规模数据和高维数据的能力，这是天文学中聚类算法必须解决的关键问题。

(3) 对聚类的结果进行准确评价，以判断是否达到最优解，这也自然要求聚类结果具有可解释性。

(4) 选取合适的聚类类别数，这是一个重要的参数。它的确定应更多地依赖于相关的经验知识以及对目标数据集所进行的必要的预处理。

(5) 对数据进行合理的预处理。该过程包括对高维数据以及对大规模数据建立索引等，它不仅是实现聚类的前提之一，也为获得更准确的聚类结果提供了一个重要的手段。

另外，聚类算法的聚类结果有一定的不可预见性，在实际应用中应根据数据类型选择合适的聚类算法，以取得最佳的聚类效果。随着聚类算法的改进和提高及广泛应用，在天文数据中发现更多的奇异天体和现象是可能的。

参考文献：

[1] Jain A K, Dubes R C. Algorithms for Clustering Data. Prentice-Hall Advanced Reference Series, 1988

[2] Jain A K, Murty M N, Flynn P J. Data clustering: A review. ACM Computing Surveys, 1999

[3] Sambasivam S, Theodosopoulos N. Advanced data clustering methods of mining Web documents. Issues in Informing Science and Information Technology, 2006

[4] http://epub.cnki.net/grid2008/detail.aspx?filename=2006054932.nh&dbname=CMFD2007, 2009

[5] 张阿品，徐保国．计算机工程与设计， 2006, 27: 384

[6] 米红娟，水 静．北京电子科技学院学报， 2007, 4: 81

[7] MacQueen J B. Some Methods for Classification and Analysis of MultiVariate Observations. In: Proc. of the fifth Berkeley Symposium on Mathematical Statistics and Probability, 1967, 1: 281

[8] Zeidat N M, Eick C F. K-medoid-style Clustering Algorithms for Supervised Summary Generation. IC-AI, 2004: 932

[9] Fred A L N, Leitão J M N. Partitional vs hierarchical clustering using a minimum grammar complexity approach. In: Proc. of the SSPR&SPR, 2000, 193

[10] 吕天阳，刘 淼，周春光，王钲旋．中国图象图形学报 A 辑， 2003, Z1: 17

[11] http://www.janoberst.com/_academics/2009_03_03_BIRCH-Efficient-Data-Clustering-Fast-Growing-Trees.pdf, 2009

[12] Guha S, Rastoji R, Shim K. Cure: an efficient clustering algorithm for large databases. In: Proc. of Elsevier Science Ltd, 2001, 1: 35

[13] Guha S, Rastoji R, Shim K. ROCK: A Robust Clustering Algorithm for Categorical Attributes. Proc. of the 15th International Conference on Data Engineering, 1999: 3

[14] Karypis G, Eui-Hong H, Kumar V. CHAMELEON: A Hierarchical Clustering Algorithm Using Dynamic Modeling. Computer Society, 1999, 32: 68

[15] http://www.paper.edu.cn/downloadpaper.php?serial_number=200711-392&type=1, 2007

[16] 吴月娴，葛临东，张 辉．信息工程大学学报， 2005, 6: 44

[17] 谢从华，朱 峰，王立军，武园园．计算机应用研究， 2007, 24: 167

[18] Ester M, Kriegel Hans-Peter, Sander J, Xu X. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. Proc. of the Second International Conference on Knowledge Discovery and Data Mining, 1996: 226
[19] Gurel L, Manyas A. Multilevel physical optics algorithm for fast solution of scattering problems involving nonuniform triangulations. Antennas and Propagation Society International Symposium, 2007, 9 : 3277
[20] Li C, Sun Z, Song Y. Lecture Notes in Computer Science, 2003, 2762: 202
[21] 陈梅兰. 计算机与现代化， 2005, 2: 1
[22] 贾世杰，黄青松，马世霞. 计算机工程， 2009, 10: 170
[23] Wang W, Yang J, Muntz R. STING: A statistical information grid approach to spatial data mining[A]. In Jarke M, Carey M J, Dittrich K R, et al. eds. Proc. Of Bases [C]. Athens: Morgan Kaufmann, 1998: 428
[24] Sheikholeslami G, Chatterjee S, Zhang A. The VLDB Journal, 2000, 8: 289
[25] http://www.math.colostate.edu/∼betten/advising/mcbee_masters.pdf, 2009
[26] 朱红灿，唐 毅. 情报杂志， 2007, 26: 101
[27] Jasmina A. ScienceDirect, 1994, 18: 3
[28] Rossi F, Conan-Guez B, Golli A E. Clustering Functional Data with the SOM Algorithm. In M. Verleysen, eds. Proc. ESANN'04, 2004: 305
[29] Guo J, Ou J, Yuan Y, Wang H. Progress in Natural Science, 2008, 18: 221
[30] 陈国平，杜利民，付跃文，王劲林. 计算机应用， 2005, 25: 2792
[31] Jain A K, Flynn P J. Image segmentation using clustering. In: Ahuja N, Bowyer K, eds. Advances in Image Understanding: A Festchrift for Azriel Rosenfeld. Piscataway: IEEE Press, 1996. 65: 83
[32] Linde Y, Buzo A, Gray R M. An Algorithm for Vector Quantization Design, IEEE Transactions on Communications, 1980, 28: 158
[33] Murty MN and Jain A K. Knowledge-based clustering scheme for collection management and retrieval of library books, Pattern recognition, 1995,28: 946
[34] Strauss M A, Weinberg D H, Lupton R H, et al. AJ, 2002,124: 1810
[35] Mähönen P, Frantti T. ApJ, 2000, 541: 261
[36] Hubble E P. ApJ, 1926, 64: 521
[37] Strateva I, Ivezić Ž, Knspp G R, et al. AJ, 2001, 122: 1861
[38] Bull N M, Loveday J, Brunner R J, et al. MNRAS, 2006, 373: 845
[39] Martĺnez V, Saar E. Statistics of the Galaxy Distribution. London: Chapman and Hall, 2002
[40] Hartigan J. Clustering Algorithm, New York: Wiley, 1975
[41] Sand D J, Treu T, Ellis R S. ApJ, 2002, 574: L129
[42] Reichart D E, Nichol R C, Castander F J, et al. ApJ, 1999, 518: 521
[43] Miller C J, Nichol R C, Batuski D J. Science, 2001, 292: 2302
[44] Paczynski B. PASP, 2000, 112: 1281
[45] Eyer L, Blake C. MNRAS, 2005, 358: 30
[46] Hojnacki S M, Kastner J H, Micela G, et al. ApJ, 2007, 659: 585
[47] Brescia M, D'Argenio B, Longo G, Pelosi S, Tagliaferri R. Earth Planetary Science and Letters. 1996,139: 33
[48] Barone F, Milano L, Russo G. ApJ, 1994, 421: 284
[49] Scargle J D. Astronomical Time Series, Maoz D, Steinberg A, Leibowitz E M. Dordrecht: Kluwer, 1997, 1
[50] Perlman E, Java A. Astronomical Data Analysis Software and Systems XII ASP Conference Series, 2003: 295
[51] Kahabka P, Li X D. A&A, 1999, 345: 117
[52] Green P D, Barker J, Cooke M P. Josifovski L. Handling missing and unreliable information in speech recognition. Proc. of AISTATS, 2001
[53] Wagstaff K L. Astronomical Data Analysis Software and Systems XIV ASP Conference Series, 2005
[54] Giavalisco M. ARAA, 2002, 40: 579

[55] Xu C, Liu J, Tang X. 2007arXiv0711.3594X
[56] Hakkila J, Roiger R J, Haglin D J, et al. AIPC, 2003, 662: 179
[57] Fuentes O, Gulati R. K. RMxAC, 2001, 10: 209
[58] Baccigalupi C, Bedini L, Burigana C, et al. MNRAS, 2000, 318: 769

The Application of Clustering Algorithms in Astronomy

YAN Tai-sheng[1,2], ZHANG Yan-xia[1], ZHAO Yong-heng[1], LI Ji[2]

(*1. Key Laboratory of Optical Astronomy, National Astronomical Observatories, Chinese Academy of Sciences, Beijing 100012, China; 2. Department of Physics, Hebei Normal University, Shijiazhuang 050016, China*)

Abstract: Clustering algorithm is an important algorithm that is used to find the data distribution and implicit scheme in data mining. It classifies large quantities of data points into several classes, minimizes the difference of the same classes, and maximizes the difference of the different classes. Especially to the large sample with multi-parameter and class unknown, the method is more brief and efficient. For better applying these algorithms, we analyze the clustering methods and their typical algorithms, present classical requests of cluster algorithms while data mining and generalize some most often used algorithms based on these requests in data mining in order to easily and quickly select a suitable clustering algorithm when facing a special issue. Moreover we summarize the theory of cluster algorithms and their application in astronomy, analyze their performance and point out their possible development trend.

Key words: astronomy; cluster algorithm; data mining