分类号_____ 密级_____

UDC_____ 编号_____

中国科学院研究生院
博士学位论文

多波段天体物理中的自动分类方法研究

张彦霞

指导教师 赵永恒 研究员

中国科学院国家天文台

申请学位级别 博士 学科专业名称 天体物理

论文提交日期 2003 年 7 月 论文答辩日期 2003 年 7 月

培养单位 中国科学院国家天文台

学位授予单位 中国科学院研究生院

答辩委员会主席 何香涛

中国科学院研究生院
博士学位论文

多波段天体物理中的自动分类方法研究

作　　者　张 彦 霞

指导教师　赵 永 恒　研 究 员

申请专业　天 体 物 理

研究方向　天 体 物 理

申请学位　博　　士

二ＯＯ三年七月

National Astronomical Observatories
Chinese Academy of Sciences

Research on Automatic Classification Methods in the Multi-wavelength Astrophysics

by
Yan-xia Zhang

Advisor: Prof. Yong-heng Zhao

National Astronomical Observatories, Chinese Academy of Sciences
Beijing 100012, P.R. China

July, 2003

摘　要

随着地面和空间观测站的建立、探测器灵敏度的提高、望远镜口径的增大和网络技术的迅速发展，天文学观测已渗透到全电磁波段，从射电、红外、可见光、紫外、X 射线和 γ 射线的各个电磁波段，形成了多波段、甚至全波段天文学，并为探索各类天体和天文现象的物理本质提供了强有力的观测手段。天文学已步入革命化的信息时代，数据以 TB，甚至 PB 计量，“数据雪崩”、“数据爆炸”已成为时代的最好概括。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，这是摆在各位天文学家面前的不得不面对的课题。在这种形势下，天文学界认为有能力且有必要建立全球性的望远镜——虚拟天文台，将全球的天文数据统一到一个实体中，为任何地方和领域的人们所利用。而且，数据挖掘技术在虚拟天文台中的成功应用，是虚拟天文台充分发挥作用的关键所在。

面对各学科、各领域的数据的冲击，数据挖掘和数据库中的知识发现从统计学、机器学习、模式识别和人工智能等学科中分流成为一门新型学科。天文学的海量数据只有借助数据挖掘技术，才能探索出隐藏在数据中的潜在的、有用的、鲜为人知的价值和信息。各学科的数据挖掘技术是天文学的数据挖掘得以发展和完善的动力和养分。由于天文数据本身的特点，有的挖掘技术可以直接借用，有的则需要调整方可使用，因而需要探索适合天文学特点的数据挖掘技术，本文正是从这方面进行探讨和研究。

如何将来自各波段的星表和图像等数据联系起来，则需要位置交叉证认来提取各种参数，从而可以研究天体在多维参数空间中的分布。基于不同天体的物理特性如形态或能谱分布在各个波段的差异，即各类天体表现特征的不同，在多维参数空间中将天体区分开是合理的、也是可行的。我们交叉证认了 ROSAT 弱源表和亮源表、2MASS 近红外数据表、USNO-A2.0 可见光数据表，同时也与一些已知天体种类的数据库 SIMBAD、VERON(2000)和 RC3 交叉证认，从而获得了多波段数据。

本文中我们提出了两种方案用来研究天体在多维参数空间中的分布。第一种方案：利用多波段数据，用自动的分类方法支持矢量机（SVM）和学习矢量量化（LVQ）对天体分类，对比了采用两个波段数据与三个波段数据的分类结果，发现随着波段的增加，分类效果越好。可见提取的天体的信息越多，越有利于天体的分类。第二种方案：针对未来天文数据维数过高的问题，我们探索了这两种方法与主分量分析方法（PCA）的混合方法，即 PCA+SVM 和 PCA+LVQ。通过主分量分析，可以知道各个参数对分类的贡献。由计算结果可知，SVM/(PCA+SVM)和 LVQ/（PCA+LVQ）方法是有效的多波段数据的分类算法。

在若干情形下，这两组方法给出的结果相当。通常前者的分类正确率高，而后者在计算速度上要快得多。这些方法可以为大规模的巡天预选源，从而提高昂贵的观测设备的观测效率。还可以将该方法用以其它种类的数据（测光、光谱、图像等数据）或这些种类数据的混合数据的分类。另外，也可从方法的研究着手，探索适合天文学的分类、聚类和离群数据的探索方法等技术。随着虚拟天文台的发展，这些方法有助于发展国际虚拟天文台的工具箱。

关键字：方法：数据挖掘和知识发现，方法：数据分析，天文数据：星表

ABSTRACT

With the establishment of ground-based and space-based observation stations, the improvement in the detector sensitivity, the increase in the telescope caliber and the rapid development of network technology, astronomical observation has covered the whole electromagnetism wavelength, from radio through infrared, optical, UV, to X-rays and γ-rays. Multi-wavelength or even the whole-wavelength astronomy comes into being, which provides strong observational means for exploring the physical essence of various objects and astronomical phenomena. Astronomy has stepped into a revolutionary information era. Astronomical data are measured by Terabyte, even Petabyte. “Data avalanche” and “data explosion” are good descriptions of the era. How to record and process the original data? How to store, combine and access data by modern computer hardware and network systems? How to rapidly and effectively explore, analysis and visualize data? Astronomers have to face all these issues. Under this situation, astronomers think that it is capable and necessary to build the global telescope——the virtual observatory, which can unify all the astronomical data into an entity, used by scientists from everywhere and every field. Moreover, the successful application of data mining technology is the key factor of sufficient exertion of the virtual observatory.

Impacted by huge data from various subjects and fields, data mining and knowledge discovery in database is developing into a new type of subject from statistics, machine learning, pattern recognition, artificial intelligence and so on. Only by means of data mining technology, we can explore the potential, useful and rare value and information hidden in data. The data mining technologies of other subjects are the push and nutrient of data mining in astronomy. Due to the characteristics of astronomical data, some data mining technologies may be directly used, others need be changed and then used. Therefore we should explore data mining technology meeting the characteristics of astronomy. In this paper, we are ready to explore and study this respect.

How to correlate catalogs and image data from different wavelengths needs positional cross-identification to extract various parameters and study the object distribution in the multidimensional parameter space. Based on the difference of physical properties of different types of objects in different wavelengths, it is reasonable and applicable to differentiate objects in the multidimensional parameter space. After positionally cross-identifying the RASS Bright and Faint Source

Catalogues, USNO-A2.0 and 2MASS released database, we cross-identified the results with Veron (2000), SIMBAD and RC3 Catalogues of known samples. Then we obtained the multi-wavelength data.

We put forward two schemes to study the distribution of various astronomical sources in the multidimensional parameter space. The first scheme is to use the two automated classification methods: support vector machines (SVM) and learning vector quantization (LVQ) to classify objects with the multi-wavelength data. Comparing the results with the data from the two bands and from the three bands, it is found that the classified results became better with the increase of bands. Obviously, the more extracted information of objects, the easier to classify objects was. The second scheme is that facing the high dimensionality of the input parameter space, the combined methods PCA+SVM and PCA+LVQ were explored. By PCA, the contribution of every parameter to classify objects was given. From the classified results, we concluded that SVM/（PCA+SVM）and LVQ/（PCA+LVQ）were effective methods to classify sources with multi-wavelength data, moreover, the two sets of methods gave comparable results in a number of situations. Generally, the former gave better results; however, the latter were considerably faster in terms of computation time. What's more, the classifiers derived by these methods can be used to preselect candidates for the large survey, and reduce time and energy wasted. Therefore the efficiency of high-cost telescope will be improved. These methods may be used for classification with other types of data, such as photometric, spectral and image data or combined data of these types. In addition, from setting about studying methods, data mining technologies of classification, clustering and outlier finding algorithms fit for astronomical characteristic may be probed. With the development of the Virtual Observatory, these methods will be useful to develop the toolkits of the International Virtual Observatory.

Keywords：Methods: data mining and knowledge discovery, Methods: data analysis, Astronomical databases：catalogues

目 录

插　图

表 格

第一章 数据挖掘和知识发现的研究及其现状

§1.1　多波段天文学

翻开人类文明史的第一页，天文学就占有显著的地位。巴比伦的泥碑、古埃及的金字塔，都是历史的见证。在中国的殷商时期留下的甲骨文里，也有着丰富的天文纪录，表明在黄河流域，天文学的起源可以追溯到殷商以前更为远古的时代。几千年来，在人类社会文明的进程中，天文学的研究范畴和天文学的概念都有了很大的发展。天文学及其分支天体物理学是以整个宇宙为对象，研究天体（包括人类赖以生存的太阳系行星系统）乃至整个宇宙的起源、结构、运动和演化。它是当今科学最具活力与基础创新力的源泉，其研究水准显示着一个国家与民族在科技发展前沿中的位置，并对一个民族的宇宙观、自然观有着深刻的影响。而且天文学和天体物理学以其研究对象的广泛性和基础性，对自然科学的众多学科有着特殊的重要意义，也是当代科学技术，特别是尖端空间技术发展的巨大推动力。二十世纪后 50 年中，随着探测器和空间技术的发展以及研究工作的深入，天文观测进一步从可见光、射电波段扩展到包括红外、紫外、X 射线和 γ 射线在内的电磁波各个波段，形成了多波段天文学，并为探索各类天体和天文现象的物理本质提供了强有力的观测手段，天文学发展到了一个全新的阶段。近几年来，以国际合作形式建造并投入使用或正在研制中的一系列大型设备，如哈勃空间望远镜，红外、X 射线空间观测站，新一代空间望远镜，以及地面巨型光学、红外望远镜和大毫米波阵等等，都反映了当今天文学和天体物理学的勃勃生机和广阔的发展前景。

光学天文学

光学天文学是利用天体在光学波段的辐射来研究天文现象的学科，是天文学中发展的最早的一部分。狭义地讲，光学天文学就是利用光学望远镜、光度测量仪器、分光仪器和偏振光测量仪器来观测和研究天体的形态、结构、化学组成和物理状态的一门学科，是实测天体物理学的重要组成部分。光学天文学是相对于射电天文学、红外天文学、紫外天文学、X 射线天文学和 γ 射线天文学而言的，因此光学天文学也是天体物理学的一个分支。

宇宙中最重要的有形物质恒星的主要辐射集中在光学波段，离人类最近的恒星——太阳使得人眼对光学波段最敏感。因而古代人用肉眼观天以定岁时；光学望远镜拓展了人类的眼界并揭示了许多新天象；先进的光学检测元件和方法使人类对宇宙的探测几乎达到了它的边缘。现代的光学天文学主要是利用大口径光学望远镜及其焦面附属仪器来研究天体的形态、结构、运动特性、物理状态、演化

阶段和化学成分的一门学科。尽管近几十年来我们发展了多种波段天文学而进入了全波段天文学时代，新发现、怪天象层出不穷，高分辨深细节耐人寻味，天文学的核心成就仍然主要来自光学天文，而且所有的新发现和新现象均要求寻找到光学对应体才能深入下去。

因此各个发达国家都在竞相独立或合作研制新一代地基或空间大口径光学/红外望远镜，如美国的口径为 10 米的 Keck I 和 Keck II 以及相应的光学干涉仪，欧洲的 16 米 VLT 和相应的干涉仪，日本的 8.2 米 SUBARU 等。正在天上的口径为 2.4 米的空间望远镜宽波段测光可以达到 30 等，角分辨率 0.01 秒，这是其他波段所无法比拟的。高光效大面积 CCD 以及大视场多目标光谱仪的出现，使得光学天文学在深度和细度上正朝着前所未有的高度发展。

下面以几个大的光学波段的巡天和数据库为例，介绍当前该波段的数据的发展情况：

DPOSS (the Digitized Palomar Observatory Sky Survey)是一个利用 48 英寸 Oschin 施密特望远镜，以 POSS-II 底片为基础进行整个北天区的可见光数字巡天的巡天计划。巡天包括三个可见光波段 g(蓝绿)、r(红)、i(近红外)（三个底片 JFN 分别记录为 Gunn 系统的 gri），覆盖了赤纬 δ=-3 度以北的天区，等效极限星等 B_{lim} 约为 22 星等。巡天结果将作为帕洛玛——诺里斯星表（the Palomar-Norris Sky Catalog，PNSC），含有 3TB 的数字信息，预计星系多于 5 千万个，恒星多于二十亿个。

SDSS (the Sloan Digital Sky Survey)是目前巡天中最富野心的巡天计划，主要集中于建立第一个 CCD 测光巡天，覆盖北银半球一万平方度，也即四分之一天区。估计该巡天星表将有一亿个源，可以为星系、类星体和恒星的最大的光谱巡天奠定基础。整个巡天计划预计历时五年，现在已经进入全方位运行阶段。其中一个 2.5 米望远镜专门用于在五个波段用扫描的方式拍摄宽视场图像（3 度×3 度）。所有的原始数据将超过 40TB，加工处理后约为 1TB，将含有一百万条光谱、一亿个天体的位置和图像参数，以及每个波段的以每个天体为中心的 mini 图像。

USNO-A2 (the United States Naval Observatory Astrometric)星表是覆盖全天的巡天表，在低于极限星等 B≈20 时含有多于 5 亿个未确定源，它们的位置可为天体测量做参考。这些源是通过 PMM (the Precision Measuring Machine)探测到的。整个巡天的全部数据量超过了 10TB。该表由一系列二进制文件组成，并且这些文件是以源的位置组织的。既然源的密度随位置的不同而不同，因此每个文件中源的数目也各不相同。要准确地提取源的参数，须提供与数据匹配的软件工具。该星表包括源的位置，即赤经和赤纬，还有每个恒星的蓝星等和红星等。

NOAO (the National Optical Astronomy Observatory)是美国的光学天文台，管

理地基的国家天文台。在国际 Gemini 项目中，NOAO 代表美国天文学界。作为国家设备，其向所有的天文学家开放，对天文学家不分地域和国界。随着巡天设备的引入和相关项目的开展，数据的积累率不断地增加，目前 NOAO 已拥有超过 10TB 的数据量。

欧南台 (the European Southern Observatory, ESO)操纵着南半球的 La Silla 和 Paranal 天文台的 4 个 8 米级的望远镜。目前欧南台数据库以每年近 20TB 的速率增长。当该项目结束时，数据量将达到几百 TB。

大天区面积多目标光纤光谱天文望远镜 (Large Sky Area Multi-Object Fiber Spectroscopic Telescope，LAMOST)于 1997 年正式立项，目前该项目正在建设中。LAMOST 是一台横卧于南北方向的中星仪式反射施密特望远镜，可观测天区的赤纬从-10 度到+90 度。相应于 5 度视场、直径为 1.75 米的焦面上放置 4000 根光纤。采用并行可控的光纤定位技术，可在较短的时间里将光纤按星表位置精确定位，并提供了光纤位置微调的可能。这将在光纤定位技术上突破目前世界上同时定位 640 根光纤的技术。通过这样的构思和设计，解决了大视场的施密特望远镜透射改正板很难做大，大口径反射望远镜视场较小的问题，使 LAMOST 成为大口径兼大视场光学望远镜的世界之最。由于它的 4 米口径，在 1.5 小时曝光时间内以 1 纳米的光谱分辨率可以观测到 20.5 等的暗弱天体的光谱；由于它相应于 5 度视场的 1.75 米焦面上可以放置数千根光纤，连接到多台光谱仪上，同时获得 4000 个天体的光谱，成为世界上光谱获取率最高的望远镜。

射电天文学

射电天文学是通过观测天体的无线电波来研究天文现象的一门学科。理论上以近代物理为基础来分析研究天体的物理特性、化学组成和结构演化。由于地球大气的阻挡，从天体来的无线电波只有波长约 1 毫米到 30 米左右的才能到达地面，迄今为止，绝大部分的射电天文研究都是在这个波段内进行的。其中类星体、脉冲星和 2.7K 微波背景的发现是射电天文对现代天体物理的三大贡献。

对于历史悠久的天文学而言，射电天文使用的是一种崭新的手段，为天文学开拓了新的园地。从前，人类只能看到天体的光学形象，而射电天文则为我们展示出天体的另一侧面——无线电形象。由于无线电波可以穿过光波通不过的尘雾，射电天文观测就能够深入到以往凭光学方法看不到的地方。银河系空间星际尘埃遮蔽的广阔世界，就是在射电天文诞生以后，才第一次为人们所认识。

为了获得高灵敏度和高角分辨率，天线的口径需相当大，最大的有美国阿勒西波的 300 米固定天线望远镜、俄国的 RATAN-600 以及德国波恩的 100 米可动天线望远镜。为了获得毫秒级的角分辨率，还用多台天线构成干涉仪及综合口径

望远镜，如甚长基线干涉仪，美国的 VLBA 等。目前正在酝酿中的接收面积达数平方公里的 SKA 和 2000 平方米的 MMA 以及空间射电望远镜干涉仪 VLBI 代表着射电仪器技术的新高峰。另一方面频谱观测技术的发展也很重要，移动频谱仪和声光频谱仪以及外差干涉仪均很重要。近几十年来，随着观测手段的不断革新，射电天文学在多个层次中发现的天体射电现象，不仅是光学天文的补充，而且常常超出原来的想象，开辟出许多新的研究领域。

现在以 NVSS、FIRST 巡天和美国的射电天文台为例，简要地介绍一下射电巡天项目：

NVSS (the National Radio Astronomical Observatory, Very Large Array, Sky Survey)是一个射电巡天项目，覆盖赤纬-40 度以北的天区。巡天星表包含超过 180 万个孤立源的全部强度和线性偏振图像测量量（Stocke 参数 I、Q 和 U），其分辨率为 45 角秒和完备极限流量为 25mJy。NVSS 的主要数据是一组 2326 个连续映射立方体（map “cubes”）,每个立方体覆盖了 4 度×4 度天区，其具有三个平面包含 Stocke I、Q 和 U 图像，还有关于这些图像的孤立源星表。每张大的图像是由大于 100 张更小的原始快照图像得到的。

FIRST 巡天 (the Faint Images of the Radio Sky at Twenty-cm Survey)是一个正在运行的射电快照巡天，覆盖南北银极近一万平方度。巡天结束时，该巡天星表将包含约一百万个源，分辨率小于 1 角秒。FIRST 巡天以牺牲观测视场为代价，获得了比 NVSS 好的空间分辨率。最后具有 1.8 角秒的射电图集是由每一个中心点附近的 12 张图像叠加而成的。该巡天的星表是从射电图集中得到的，包括峰值流量、积分流量密度和由拟合二维的高斯图像得到的源的大小。近 50%的源在 POSS I 底片的极限内（E≈20）具有可见光对应体。在小于 5%的错误率下，V 达到 24 的源都可以光学证认。选择的巡天区域尽量与 SDSS 巡天一致。在 SDSS 巡天的极限星等范围内，FIRST 巡天中的 50%的源探测到了光学对应体。

美国的射电天文台(the National Radio Astronomy Observatory，NRAO)建于 1956 年。其拥有的数据量也不下 10TB，可以提供若干巡天机会如 NVSS 和 FIRST 巡天以及 GBS 巡天(the Green Bank Survey)。

红外天文学

红外天文学是利用电磁波的红外波段研究天体的一门学科。整个红外波段，包括波长 0.7～1000 微米(1 毫米)的范围。通常分为两个区：0.7～25 微米的近红外区和 25～1000 微米的远红外区；也有人分为三个区：近红外区(0.7～3 微米)、中红外区(3～30 微米)和远红外区(30～1000 微米)。

随着半导体物理学的发展和军事侦察的需要，研制出了灵敏度很高而热噪声

很低的单元（测辐射热计）和阵列红外检测器件（红外 CCD），红外天文学在近年获得了巨大的发展。已经和正在研制的大口径光学望远镜均是与红外共用的。当然不仅红外检测器本身的热辐射会妨碍对微弱信号的检测，天空背境和环境的热辐射也是讨厌的噪声源。因此红外检测元件和一些核心的相关部件必须在液氮或甚至液氦条件下工作。特别是中红外和远红外，需要到地球大气外去工作。迄今最重要和最成功的红外探测计划是口径 60 厘米的 IRAS 红外天文卫星（1983 年发射，观测到 245839 个红外源）。其次有 ISO 中红外空间天文台，大视场红外实验装置和深空近红外巡天装置等。宇宙背景探测器（COBE）也包含了红外波段，对 2.74K 背景辐射的探测起了巨大的作用。红外波段对于研究星系的起源和恒星及其行星系统的起源是十分重要和有用的。因此美国计划发射空间红外望远镜装置（SIRTF），同温层红外天文台（SOFIA），并在地面建造口径 8 米的红外专用望远镜（IRO）等。

下面以 2MASS 为例介绍红外巡天，以 IRSA 为例介绍红外数据库：

2MASS (the Two Micron All-Sky Survey)是一个近红外（J、H 和 K_s）的全天巡天项目。该项目始于 1997 年，数据于 2002 年度释放完毕。2MASS 利用了两台新的、高度自动的 1.3 米望远镜，每台配备一个具有三个通道的照相机，能够同时在三个波段 J（1.25μm）、H（1.65μm）和 K_s（2.17μm）观测天空。当巡天结束时，2MASS 星表包含近 3 亿颗恒星、50 万星系和星云在三个波段的天体测量和测光属性，以及多于 12TB 的图像数据。

IRSA (the NASA Infrared Processing and Analysis Center Infrared Science Archive)提供了获得各种数据的服务，主要针对红外数据。IRSA 也提供了交叉证认工具，用户可以根据要求从各种源中选取候选目标。IRSA 主要提供以浏览为基础的查询服务，包括星表和图像。IRSA 含有多于 15TB 的数据，大部分与 2MASS 巡天相关。

X 射线天文学

X 射线天文学是通过 X 射线波段(波长 0.01～100 埃的电磁辐射)研究天体的一门学科。因为天体的 X 射线会受到地球大气的严重阻碍，所以主要利用卫星进行探测。因此，虽然 X 射线的探测始于二十世纪四十年代，但是成为一门学科，则是在人造地球卫星上天以后。

X 射线天文学从诞生时起，在近二十年的短暂时间内发现了一系列前所未知的新型天体，获得光学天文和射电天文所无法得到的天体信息，大大地扩展了天文学的研究领域。X 射线天文学所显示的独特威力，使得它在当代空间天文学中处于特别重要的地位。钱德拉 X 射线天文台是 NASA 的第三大天文台，提供了

X 射线波段的高分辨率的图像和光谱数据。钱德拉 X 射线天文台是 1999 年 7 月份由哥伦比亚号航天飞机发射升空的。钱德拉数据库（the Chandra Data Archive，CDA）是史密松天体物理台管辖的钱德拉 X 射线天文台科学中心（the Chandra X-ray Observatory Science Center，CXC）的组成部分。由于 HEASARC 数据库研究中心是与 X 射线波段相关的空间数据库，故钱德拉 X 射线数据也存放在该数据库研究中心。钱德拉数据库包括科学相关数据和工程数据两部分。另外，著名的 X 射线观测卫星有自由号（1971 年发射，全天探测共得 340 个源）、爱因斯坦号（1978 年发射，灵敏度和角分辨本领均高了一个量级，探测到 4809 个源）、EXOSAT(欧洲 X 射线天文台卫星，1983 年发射)、ROSAT（伦琴号卫星，1990 年 6 月 1 日发射，工作于 0.1-3 电子伏特，发现了 60,000 多个 X 射线源，包括 25000 多个活动星系核、20000 多个恒星和 5000 多个星系团）。宇宙学和天体物理学高级卫星（飞鸟号，1993 年 2 月 20 日发射）、日本人发射的白鸟号和阳光号卫星等。卫星上搭载的仪器为掠射式望远镜和高能探测器。早期研究集中于太阳，弄清了它的宁静、缓变和突变三种辐射成分，后来逐步扩大到各类天体。既有点源也有面源；既有 X 射线星也有 X 射线星系和星云。X 射线星中最重要的一类为激变双星。它的一个子星为密度极高的恒星残核，贪婪地吸积着膨胀的老年恒星流出等位面的物质并在自身周围形成一个吸积盘。在盘上的某些地方出现热斑并产生热核反应而辐射 X 射线。弥漫 X 射线星云多为超新星遗迹。X 射线星系多为活动星系核，它们也是吸积过程的产物。星系团也有很强的 X 射线辐射。或者源于弥漫星际介质，或者孤立点源，与团成员的种类和内涵有关。

紫外天文学

利用天体在 100 到 4000 埃的紫外波长的辐射来研究天文现象的学科。由于大气对紫外波段的吸收十分严重，因此需要到高空或大气外进行观测。由于氢原子赖曼线系限外的连续吸收以及与光学天文学的交叉，紫外天文学的研究范围实际上只限于 912～3000 埃之间。由于元素的中性和电离态的共振线在紫外区比在可见光区丰富得多，共振线对研究天体的物理状态和化学组成极为敏感，因此我们很有必要把观测波段扩大到紫外区。当然第一个研究对象是太阳，对研究色球和日冕间过度层以及耀斑活动提供了有价值的信息。对太阳系内的行星和彗星等天体的紫外光谱、反照率和散射的观测，有利于确定它们大气组成，从而建立大气模型。

1978 年 1 月 28 日发射的 IUE 地球同步卫星载有一架口径 45 厘米的卡塞格林望远镜和两台摄谱仪（高色散和低色散），工作于 1150-4000 埃之间。发现了大量的紫外天体并编辑出版了 IUE 星表。在 1990 年 6 月 1 日发射的 ROSAT 卫

星上还载有 EUV（极端紫外）望远镜，探测能量在 25-100 电子伏特间的源，结果共发现 384 个源，其中主要为白矮星和晚型活动星，其他为激变变星和河外天体。1992 年 6 月 7 日发射了 EUVE（极紫外探测者）卫星，上载三个掠射扫描望远镜和一个谱望远镜（50-740 埃）。发现的天体中 55%为晚型星，30%为白矮星，其他为激变变星、早型星和河外天体，其中最亮的源为 εCMa，一个光谱型为 B2II 的蓝巨星。目前正在天上工作的哈勃空间望远镜也有紫外观测仪器，是这一领域中的最大者。

γ 射线天文学

利用波长短于 0.01 埃的辐射来研究天文现象的学科。γ 射线被地球大气严重吸收，因此只能利用高空气球、火箭和卫星搭载仪器进行观测。能量高于千亿电子伏特的甚高能 γ 射线穿过地球大气时会产生高能粒子簇射，从而形成切仑科夫辐射。可以用特殊的大口径望远镜来探测这种辐射而间接探测到 γ 射线源。中国科学院高能物理研究所和北京天文台合作在北京天文台兴隆观测站建了两套这样的甚高能 γ 射线望远镜。每套望远镜由三个口径 1.5 米的聚光镜同轴组成，每个聚光镜的焦面上有一个光电倍增管。只有三个倍增管在微秒级的时段内同时接收到信号才算是探测到了有用的信号。进一步还可建造成像式切仑科夫望远镜。美国霍普金斯山的拼镜 10 米惠普切仑科夫望远镜就是一例。目前还正在研制多台这样的望远镜。搭载在卫星上的 γ 射线望远镜则是利用 HADAMA 编码孔板来鉴别源的方向，其检测器多为高能粒子闪烁计数器，可根据不同能域而选择不同种类的闪烁体。

1958 年莫里森从理论上预言某些天体可能发射强的 γ 射线。接着发现了太阳的 γ 射线爆发，它总是伴随着射电爆发。1962 年月球轨道卫星“徘徊者”3 号和 5 号发现了宇宙 γ 射线背景辐射。这一发现为后来的轨道太阳观测台 3 号、阿波罗 15 号和小型天文卫星 B 等所证实。1975 年 9 月发射的 COS-B γ 射线卫星的资料表明大部分 γ 射线来自银河系内，似乎起源于高能宇宙线与星际介质的碰撞。第一个全天探测 γ 射线望远镜 CGRO 于 1991 年 4 月 5 日发射，带有 4 个科学仪器(爆发和短暂出现源探测仪、有方向性闪烁能谱仪、成像望远镜和高能 γ 射线实验望远镜)，重 15.9 吨，轨道高 455 千米，角分辨率 2 分，时间分辨率 0.1 毫秒。探测得知 γ 射线爆是各向同性的，但在空间分布上是不均匀的。因此既不能肯定它们是河外天体也不能肯定它们是河内天体。此外还发现了 γ 射线活动星系核和新型高能河外 γ 射线源等。INTEGRAL（国际 γ 射线天文台）工作于 15 千电子伏特到 10 兆电子伏特间，搭载锗能谱仪和碘化铯成像器，探测器面积为 2500 平方厘米，能量分辨率在兆电子伏特处为 500，角分辨率半强度宽度为 17 分。

并配有同时在 4-100 千电子伏特记录 X 射线和在 5500-8500 埃记录可见光像的仪器，以期通过光学证认来弄清 γ 射线爆的本质并发现更多的新现象。

多波段数据库

正是基于各种探测仪器的发展和地面与空间观测站的建立，多波段天文学也随之发展和壮大，使得多波段数据的获得成为可能，从而我们可以研究天体的多波段特性。这里介绍几个多波段的数据库和数据服务平台：

MAST (the Multimission Archive at the Space Telescope Science Institute)包含有各种天文数据，主要来自可见光、紫外和近红外波段。MAST 提供了交叉证认的工具和独立任务的查询能力，而且可以预览图像和光谱，便于存档用户校对。MAST 的主要数据来自哈勃空间望远镜，目前总的数据量已超过了 10TB。

高能天体物理科学数据库研究中心（the High Energy Astrophysics Science Archive Research Center，HEASARC）是一个多任务天文数据库，侧重于极端紫外、X 射线和 γ 射线波段数据。HEASARC 目前拥有来自 20 多台天文台的 30 年的 X 射线和 γ 射线的观测数据。目前的数据量已经超过 5TB，随着当前或不久后大量的高能物理卫星的发射，该中心的数据量会以更快的速度增长。

SkyView 是 HEASARC 开发的基于网络的平台。用户可以通过 SkyView 获取任何天区从射电到 γ 射线各个波段的图像数据。SkyView 最强大的功能是能够处理各种几何转换和坐标变换以满足用户对数据格式的要求。

国家空间科学数据中心（the National Space Science Data Center，NSSDC）向用户提供了基于网络和离线的方式获取来自 NASA 项目的各种数据，其中包括每年来自宇宙微波背景探测器的几个 TB 数据量。NSSDC 是戈达德飞行中心在 1966 年建立的第一个数据中心，并且还在不断地收集和整理数据，包括独立的和非独立的数据。NSSDC 通常作为所有 NASA 空间项目的数据的最终存储库。

行星数据系统（The Planetary Data System，PDS）包括来自 NASA 行星计划的数据、天文台观测数据和实验数据。PDS 提供了获取来自 Pioneer、Voyager、Mariner、Magellan、NEAR 太空计划的数据的服务，以及其他有关陨石、彗星和行星（包括地球）的科学数据的服务。

总结

综上所述，各种空间观测设备的投入运转，以及一些大型地面观测手段和新技术的应用，使得多波段天文学正处在一个蓬勃发展的新时期，已经并将继续取得一系列激动人心的发现。来自各个波段的数据量以指数量级增长，以 TB 甚至 PB 计量。但是我们不可因获得如此巨大的数据量而沾沾自喜，如果不能有效地

认真地加工、处理和分析数据，那么我们只能面对数据海洋，望洋兴叹。如何有效地科学地探索来自数字巡天和数据库的新的好几个 TB 的数据？如何从具有几十亿甚至几百亿的天体或如此大的数据变量的数组中进行科学发现？这是摆在天文学家面前不可回避的问题。为了有效地处理这些问题，天文学家们决定建立全球性的虚拟天文台，以应对形势发展的需要。发展适合天文发展和需要的数据挖掘与知识发现技术，充分有效地从数据矿山中挖掘出天文学家感兴趣的和有意义的天体和天文现象，从而推动天文学理论的进一步发展和完善。

§1.2 虚拟天文台

随着望远镜、探测器、计算机和互连网技术的发展以及大量天文数据和资源的网络共享，天文学界认为有能力且有必要建立全球性的望远镜——虚拟天文台，将全球的天文数据统一到一个实体中，为任何地方和领域的人们所利用。虚拟天文台的出现和发展，预示其在二十一世纪将具有广泛的应用前景，将对天文学的发展起到巨大的推动作用，并在知识和技术等方面对天文学家提出了新的挑战。而且，数据挖掘技术在虚拟天文台中的成功应用，是虚拟天文台充分发挥作用的关键所在。

虚拟天文台的兴起

200 多年来，天文学研究的方式一般是由单个天文学家或几个天文小组对数量相对较少的天体进行研究。这样，为了获取足够多的数据，整个天文学界花费了大量的时间和精力，但得到的数据仍不足以进行统计研究。而且由于使用大型观测设备的时间很有限，使得许多需要大量高精度观测数据的天体物理问题没有得到解决。

但是，近十年来，各种技术正在经历史无前例的飞速发展，使得天文研究的方式、方法发生了巨大的变化。望远镜口径的增大、大面积探测阵列的发展、计算机运算能力的迅速提高和通讯网络的不断普及，都给天文学带来革命性的变化：可以建造新一代大口径地面可见光和红外望远镜、以及毫米和厘米波段单天线阵和多元天线阵；一系列天文学新分支的兴起，如 γ 射线天文学、X 射线天文学和红外天文学；许多正在酝酿或计划中的地面天文台和空间天文台大天区的数字巡天计划覆盖了较大的波长范围，从 X 射线、紫外、可见光、近红外到微波、射电波段[1-2]；随着大的巡天计划的实施以及大型天文观测设备的建造，将产生大量的数据，通常以 TB（10^{12} 字节），甚至 PB（10^{15} 字节）计量。更多的地面和空间天文设备、以及更大口径和更精密仪器的投入使用，将带来天文数据的进一步飞速增长，例如哈勃空间望远镜每天大约产生 5GB 的数据，筹建中的大口径综合巡天望远镜(Large-Aperture Synoptic Survey)日产数据将高达 10TB。因此，Szalay 认为天文学正面临着一场“数据雪崩”[2]。

而且，未来几年天文数据仍在迅猛地增长，其广度和深度是以往数据所不可比及的，将为从事不同科学研究项目的科学家提供良好的研究素材。为了使用的方便，将数据系统列表化是非常重要的，这可以减少对以前观测过的源的重复观测。拥有了百万甚至上亿个天体的多波段数据，使得天文学家可以对其进行分析，同时要求具备与之匹配的分析工具，并为数据挖掘提供用武之地，例如复杂形态的证认、大样本的交叉证认、发现稀有天体和一些天体的时间演化序列。具有如

此大的数据量，在天文学历史上第一次可以将复杂的数字模拟与统计多变量分析进行比较，结束了数值模拟结果因缺乏足够的观测数据而难以验证的历史。

科学技术的突飞猛进为古老的天文学注入了新的活力，多波段、10^{12}字节的数据联合在一起不再是异想天开。在天文发展的数字化时代，人们逐渐意识到获得数据、组织数据、分析数据、传输数据，是科技持续增长的基本要素。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[3]。这些因素决定有必要建立一个机构，使其能够充分地有效地实现这些技术，因此，这就需要建立类似虚拟天文台的机构来管理这些正在增长的天文数据。很快人们就会意识到投入巨大的财力物力将所有的数据联合到虚拟天文台中，远比建立一个传统的天文台有意义得多。这样，天文学家可以很容易地迅速对特定天区进行观测，而不是坐等几个月或一年才可获得望远镜的观测时间。由此可见，数据库的联合是大势所趋[2]，虚拟天文台的建立是必然的[4]。

在这种形势下，美国国家科学院在天文与天体物理发展的新十年展望中把国家虚拟天文台列为第一优先发展项目。国家虚拟天文台（the National Virtual Observatory，简称 NVO）的概念一经提出，立即引起天文学界的广泛重视。各国也纷纷相继出台了自己的虚拟天文台计划，如英国的天文网格项目（UK Astro-grid）、欧洲天体物理虚拟天文台（AVO）、俄罗斯虚拟天文台、澳大利亚虚拟天文台、印度虚拟天文台、加拿大虚拟天文台、意大利虚拟天文台，另外我们也不甘落后提出了中国的虚拟天文台。并且以美国、英国、欧洲为首的虚拟天文台组织在 2002 年 6 月在德国的“国际虚拟天文台大会”上决定成立国际虚拟天文台联盟（IVOA）。因此，虚拟天文台的建立可以说是信息时代发展的产物。

天文大型数据库

当代天文学的主要特征是走向全波段的观测和研究，在此过程中重要的一步是空间天文学的兴起。巡天技术的发展使得天文学和空间科学正在发生着巨大的变革。随着大量的遍布各国的地面观测站（如美国可见光天文台 NOAO、美国射电天文台 NRAO 和欧洲南方天文台 ESO）和空间观测站（如哈勃空间望远镜 HST）的飞速发展，数据组正在如潮水般的涌来，而且巡天项目在规模和数量上仍在持续增长。

目前有大量的分布在不同场所的不同数据系统的存档数据，如 ADS、SIMBAD、NED 等。在地面天文和空间天文方面已有一批来自数字巡天和特殊的巡天计划的观测资料。例如：

红外：IRAS、2MASS、DENIS、ISO、SIRTF…

光学：SDSS、DPOSS、USNO-A2、MAST、POSS、UKSTU、ESO 底片巡天、LCRS、EIS、CfA/ZCAT、2dF、GSC-II、ROTSE…

射电：FIRST、NVSS、WENCS、SUMSS、GB6、PMN、nC…

紫外：IUE、UIT、OAO-3、UIT、WUPPE、HUT、FUSE、ORFEUS、BEFS…

X 射线：EXOSAT、XMM-Newton、CDA、ROSAT…

γ 射线：CGRO…

多波段：HEASARC、MAST…

微波背景辐射：COBE、MAP（以及将来的 Plank 卫星）

其中如 2dF、SDSS、2MASS、DENIS、FIRST、Chandra、XMM 和 MAP 正在运行，一些新设施如 VISTA、LSST、CELT、NGST、VST、UKIRT WF、SCUBA-2、SALT、ALMA 和 Plank 将在近几年内启动。随着巡天产生几十 TB 甚至几百 TB 的天文数据，数据的存储、分析是摆在当前天文学家面前的紧迫任务。天文学家必须借助虚拟天文台才能将各个不同波段的数据统一成一个整体，充分地、深入地进行研究和探索。

虚拟天文台的主要功能

虚拟天文台的主要功能[2]是把分散的数据联合起来。数据存放在各个熟悉该数据的课题组那儿，即数据分散在各地。各课题组应保留和改进自己的数据，提供表示各数据库的内容、文件和元数据。

就使用数据而言，各种用户对数据的要求各不相同。大多数普通用户仅是粗糙地浏览，用 WWW 界面作大量的十分简单的查询。对中心网站来说，支持这些不太复杂的查询引擎是相当容易的。中等水平的用户要较详细地使用数据库。而当高级用户对 10^{12} 字节的数据多次查询并提取 10^{9} 字节的数据去进一步研究时，将遇到很多困难。大多数查询各具特色，至少在一段时间内不会完全一样。通常，科学家开始在有限的数据范围内探索天体的多波段特性，逐渐转向更复杂的查询。一些重要的工作需要软件支持，例如在不同的数据库中浏览同一天区的天体，探索其独立特性，用一些条件限制在主要天区，并以在天区上天体间的角分辨率为基础进行寻找，与其它数据库的数据进行交叉证认，创建个人数据组和新的数据库。

而且，数据就其本身而言是多维的，每一个天体可以通过流量、在天区的位置、尺度大小、红移等来表示。寻找特殊类型的天体（如类星体），须在 N 维空间中定义复杂的区域。需要考查空间关系，如找最近的天体，或找满足一定角距离的其他天体。如果满足标准的天体的数量太多，那么快速地获取数据的文件就

建立不起来。唯一的办法是与数据库本身连接，直接将数据传送给分析工具。

改变天文研究特征的主要技术是宽带高速的信息网络传输技术。高速与广泛分布的信息网络的快速发展，意味着可以使世界各国的天文学家获得这些数据。未来的网络传输速度将是相当快的，从而能从地面设备和空间设备中高效率地获取数据，并将数据高效率地传输到不同的场所。其科学潜力是不可估量的。利用下一代高速网络，可以实现不同网站间的相互联系。虚拟天文台将成为创新网络应用的主要范例。

联合机构的创建和维护包括几方面内容。首先要了解一些基本要求和定义适当的标准，然后建立满足这样条件的工具。新的分析工具、新的存挡方法及与之相应的软硬件的创建需要通过提高计算机科学技术来实现，这远远超出天文学的范畴。现在的社会是高度的信息反馈化社会，这些技术的发展仅有天文学家的参与是远远不够的，必须与计算机科学家、统计学家、甚至来自信息工业的参与者广泛合作。

虚拟天文台的科学目标

⑴ 多维观测参数空间的探索

虚拟天文台的主要科学目标之一是探索多波段巡天测量得到的源的多参数空间。这样的工作已经做过，每个源在多参数空间中被当作一个点或一个矢量。但是将各个巡天数据统一到虚拟天文台中，将会有更广泛的复杂的应用。这些数据库将提供全天或几乎全天在 13 个不同波段的信息，在多维空间里展示全天的真面貌。每一个数据库的每个波段包含 10 亿个天体，多达好几个 TB。必然存在不同数据库间的重叠，同时在两个不同数据库间仍有许多未知的事件。一旦数据库完备，它们的联合将只允许一对一的查询。在不久的将来，可以设想用图像或像素、或用数据和图像来进行探索，而且可以在巡天中考虑各天体和天文现象随时间的演化，从而可以进行时间演化序列的探索。例如探索纯数据组时充分利用数据的丰富信息（成百上千的测量参数），从较平常的类型中区别出有趣的天体，这很有可能发现以前未知的天体和现象。更详细的评述可参看 Djorgovski 等人的文章[5]。

通常，一个完备的观测参数空间包含的量有天体的坐标、速度或红移，有时也包括运动方式、一定波长范围内的流量（光谱、一组波长的图像）、表面亮度、确定源图像的形态参数、一定时间范围内的能谱变化参数等等。任何巡天都受到自身的选择效应和测量极限的限制，只能测得一部分观测参数，例如流量、表面亮度、角分辨率、光谱分辨率、样本大小等等。这样任何巡天数据提供的宇宙图像都是有限的，因而人们对宇宙的了解也只能是有限的。但是，多个巡天数据在

虚拟天文台中的完美结合将克服这些限制，得到更加完善的真实的宇宙图像（层次化的、大范围的、概括性的等等）。

有时多波段交叉证认可以发现有趣的天体和现象，如类星体和强射电星系的发现，各种 X 射线源和极亮源的发现，通过研究余辉在 γ 暴研究方面的进展等等。利用多波段星表和图像以及源的性质与环境来搜寻未知类型天体，或搜寻稀有天体（如高红移类星体和褐矮星）；其它还有用 DPOSS 的 JFN 三色图像搜寻低表面亮度星系和光学星系团；用 XMM/Megacam-VST/VIRMOS 寻找 z>1 的类星体和星系团；使用 ROSAT 和 DSS 进行的 AERQS 类星体巡天和搜索年轻星及褐矮星；使用 ROSAT 和射电寻找 Blazar 的 DXRBS 巡天等，数字化历史底片研究恒星的长期变化，多色图像巡天寻找高红移天体的 2dF 类星体巡天、CADIS 和 EIS 巡天等。对虚拟天文台而言，与以前知之甚少的观测参数空间相比，多波段研究将会对天体性质的了解更加深入，从而进一步了解天体的多波段特性。

二十世纪三十年代，Zwicky[6]开创性地提出通过系统地研究观测参数空间来探索宇宙空间的概念，遗憾的是当时赞成他的想法的人并不多，又由于当时观测技术的限制，他的想法未得以实现，否则他将是发展和利用虚拟天文台的第一人。但这个富有创意的思想在 1975 年 Harwit[7]和 1986 年 Harwit 与 Hildebrand[8]加以更深辟的讨论，他们认为开辟新的观测参数空间对探索新的发现很有意义。随着大量的巡天数据的网络共享、探索和分析技术的提高，人们正准备将这一想法付诸实施。

⑵ 稀有天体与新型天体的发现[3, 5]

寻找稀有的已知类型天体（如高红移类星体、褐矮星）的巡天项目正在蓬勃发展。稀有性往往是观测的选择效应造成的，就褐矮星而言，在宇宙中是很普遍的，只是难以发现而已。已知类型天体的性质可能与巡天仪器的参数混在一起，如带通、流量等，在参数空间的特殊区域中，满足一定选择标准的天体才能被发现，例如寻找高红移类星体[9-13]。类似的技术同样适用于寻找高红移星系[14-15]和褐矮星[16-20]。

当然，在多参数空间中发现更多未知的稀有天体或现象更有意义。例如可以在大样本中寻找一些在小样本中不可能发现的罕见事件：假如某种有趣的天体出现的概率是百万分之一或一亿分之一，那么需要几百万或几亿个样本才有可能发现。稀有天体在某些参数空间中辨别不出来，但是在其它一些参数空间中却可以区分出来。在可见光和红外波段象恒星的未确定源，仅有的区别是宽波段的能谱分布，而能谱分布是以不同带通间的流量比例为参量的。在色参数空间中搜寻到奇异天体，还要对候选者进行光谱证认才可以确定其类型。1998 年 Djorgovski 等[21]在 DPOSS 巡天中利用色参数空间发现了高红移类星体和 II 型类星体。在色

参数空间中，正常恒星分布呈现香蕉形状，并形成一个温度序列，而类星体具有不同于正常恒星的颜色，远离恒星区，从而以在色空间的不同位置就可以将类星体与恒星区分开，如图 1.1 所示[3, 5]。然后再根据吸收线和发射线的特点就可以区分开高红移类星体和 II 型类星体。这两类类星体相当稀少，面密度小于每平方度 10^{-2}，低于可靠的恒星与星系分类的界限，这样，为了统计性地探测一些有意义的样本，就需要大天区巡天并且依靠合适的选择方法。同样的方法也适合于其它波段的低角分辨率的巡天，例如在 IRAS 数据中区分恒星和星系，用射电指数把类星体从射电星系中辨别出来，用 X 射线的硬度比从样本中找出 AGN 等等。同样，在可见光和近红外波段可以找到各个红移的类星体的完备样本[22-23]；可以选出特殊谱类型的恒星用以探测银河系结构[24]；如果星系的形态可以参数化，则可以对特殊类型的星系分类并将其挑选出来[25]。

相比之下，在大的数据组中寻找稀有的未知类型天体将具有更加诱人的前景。这可以通过系统地寻找参数空间中的不同于大多数天体分布者来发现，即从统计学的角度显然不同于其它天体。SDSS 和 DPOSS 研究组在寻找高红移类星体时就利用这种方法发现了一些新类型的天体。这些天体具有不同寻常的，甚至目前不太了解或根本不清楚的光谱，这些光谱使它们具有特殊宽的颜色，巧合地分布在色空间的高红移类星体分布的区域。从而在未探索的参数空间中系统地寻找离群数据可以发现一些特殊天体，其中一些结果便是新天文现象的原型。如果这些新的天体或天文现象确实存在，而且可以在已有的数据中探测到，那么在大范围内进行彻底地无偏差的多波段的宇宙探索将可以发现它们。可见，虚拟天文台可以促进新发现。

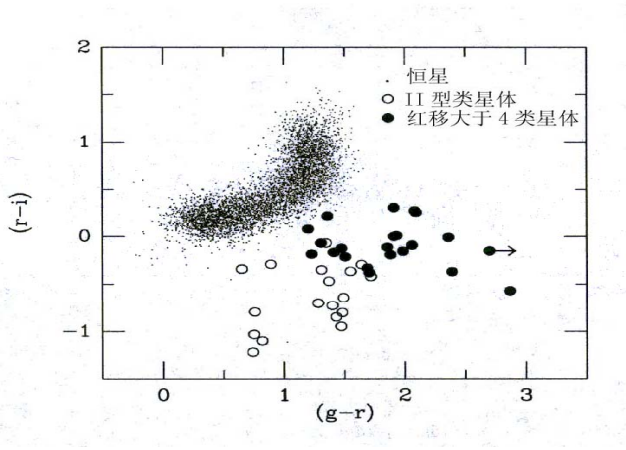


图 1.1　DPOSS 巡天得到的三色空间中天体的分布，其中点代表 r 约为 19 的正常恒星，实心圆圈代表 z>4 的类星体，空心圆圈代表 II 型类星体

⑶ 边缘学科的兴起

尽管虚拟天文台对发现稀有天体具有较大的科学潜力，但更重要的是其对任何要求融合数据来研究天文现象的项目，都有着不可忽视的影响。虚拟天文台的出现促进了多波段天文学的发展，不同波段的巡天资料的联合可以从更深层次上探索宇宙。不过，大多数工作最终要在光学波段证认，例如射电源或 X 射线源的光学波段证认。对软件的基本要求是准确有效地从多波段交叉证认探测源。如果是一对一，那么工作就容易多了，但如果是一对多，则需要更复杂的算法和利用一些背景知识进行优化。同时，虚拟天文台推动了各种令人兴奋的科学探索，如活动星系核和星系团的多层次研究、低表面亮度星系的形成和演化的研究[26-28]及星系结构的研究，这些研究对观测提出了挑战，这就需要充分发挥虚拟天文台的内在潜力。

虚拟天文台的出现也促进了统计天文学的兴起[1, 3]，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、AGN、星系团等）完备样本的建立与研究。天文学研究的热点和难点是如何系统地、统计地定量分析宇宙与将宇宙图像化及发展用于支持限制理论模型与理论思维的方法。虚拟天文台的建立可以使科学研究在数量和质量上充分地提高，从而不再受小样本泊松误差的限制。但是，了解巡天数据的系统误差和偏差将越来越重要，而且对这样的研究，样品的个数和大的视场是很重要的。从某种意义上讲，我们正在对天体的种类进行直接探索，通过新的数字巡天获得的信息将使之达到更高的准确度和更深层次的细节。

数据挖掘技术

在数据化、信息化的今天，数据挖掘应运而生并成为一种新型学科。数据挖掘（DM）是指半自动或自动地从海量数据中发现模式、相关性、变化、反常规律性、统计上的重要结构和事件。在天文上, 就是从海量数据中发现稀有的天体或现象, 或者发现以前未知种类的天体或新天文现象。不管天体是已知的或未知的，数据被划分成各种不同类型的天体时，将遇到自动分类或聚类分析的问题。这是正在快速发展的数据挖掘（DM）和知识发现（KDD）领域的一部分。有关这方面的课题和方法可参看 Fayyad 等人的文章[29]。

虚拟天文台中典型的数据组具有一些性质：在 100 维中约有 10^9 数据矢量，数据组由若干种类的天体组成，如不同光谱类型的恒星、不同哈勃分类或形态的星系、类星体等等，这必须依靠强有力的分析技术支持。

通常天文学中使用的数据挖掘技术有：

① 监督的分类方法，如人工神经网络（ANN）或决策树。这种方法通常用

于区分恒星与星系[30-32]，在多参数空间中寻找具有预测特性的已知类型天体也可以用这种方法（如寻找高红移类星体）。

② 非监督的分类方法[34-37]，如 EM(Expectation Maximization)，MCCV(Monte Carlo Cross Validation)。这些方法已用于确定数字巡天得到的星团数目，并将成为虚拟天文台分类工具的重要组成部分。

③ 主分量分析方法（PCA）[38-40]，具有非监督性，对数据进行预处理，去掉一些无关或不重要的参量，即降维。主要用于恒星、星系和类星体的光谱分类，星系的形态分类。

④ 其它方法，如最大似然法、非参数技术、信息瓶颈、小波、广义 Hough 变换、贝叶斯方法、独立分量分析方法（ICA）、最近邻规则、最小距离方法等。

一旦数据被分成明显的种类，就可以对它们进行解释。问题是：某一类天体的性质间是否存在有趣的相关性？这些相关性可以反映新的天体物理知识，例如恒星主序、星系基平面的相关性、Tully-Fisher 关系等，但同时这也使得对聚类分析的统计解释复杂化。怎样才能证认出相关量并区分出无用的观测量？考虑到这些问题，在现实中盲目地运用聚类分析方法会产生误导或错误的结论。聚类分析方法必须足够有效地解决这些问题，而且分析结果必须有可靠的统计理论基础。聚类分析的优点在于有助于划分数据空间，从而找到不同寻常的天体。应用在多参数空间中的数据可视化技术是聚类分析的另一重要组成部分，好的可视化技术对解释观测结果是很重要的，尤其在重叠的情况下。另一重要课题是在处理虚拟天文台中的大量巡天数据时提出的各种问题，一些算法和模型需要相互作用或反复使用，甚至需要寻找新的算法。

总之，数据挖掘技术是虚拟天文台的重要组成部分之一，该技术在虚拟天文台中成功的应用，能使任何地方的科学家和学生在不依赖于大望远镜的情况下就可以做出一流的工作。应用这些方法可以有效地处理天文学中的“数据雪崩”，发现新类型的天体、星团，并从结果中得出一些新的有意义的天体物理知识，这对天文学发展是尤为重要的。

虚拟天文台的结构及其设计原则

当前天文学中重要的一项任务就是如何将高度分散的巡天数据联合起来推动知识发现，能胜任这项任务的正是虚拟天文台，其结构如图 1.2 所示。Brunner 以数据服务的难易程度将研究项目分为两类：基础服务和高级服务[40]。基础服务中的数据已经作为商业数据的一部分，并且大多数数据库中心已经使用。基础服务包括星表搜索引擎、系统元数据、数据关联、图像元数据引擎、图像数据库获得及查询优化等。高级服务远超出平常的服务，要求对处理过的数据再加工才能

应用，这样必须用特殊的软硬件支持才行。高级服务包括计算与数据网格、图像处理、模式识别、统计分析、可视化、机器学习等。Brunner 倡导的数据库设计原则是不论在实际上还是应用上都要封装数据库，这样可以省去提供不同数据库间相互作用的努力（如即插即用模型）。这种方法要求提供相应的方案给数据库，从而可以在当地的通信水平上充分发挥已有设备的作用。这种服务方式不仅减少了在通信上的投资，而且促进了提供分析工具者之间的协作。

虚拟天文台是高科技发展的产物，显然不同于与传统概念上的天文台。其设计原则：第一，必须是发展的，随着数据量的迅速增大、计算机网络的快速发展、科学的进步和用户的要求，虚拟天文台的各种软硬件设施都要及时更新换代以保持其先进性；第二，是分布式的，无论其数据还是各种计算机软硬件资源都将分布在全球不同的国家和地区；第三，是完整的统一体，尽管地理位置上具有分布性，但功能上是有机的统一体，对用户提供整体的服务；第四，提供公众服务，为公众了解天文，学习和利用天文提供服务；第五，是面向全球的，它的资源将为全球天文学家所共享。此外，虚拟天文台开创了通向未来的大道，为天文学研究提供了盛况空前的前景和视角。

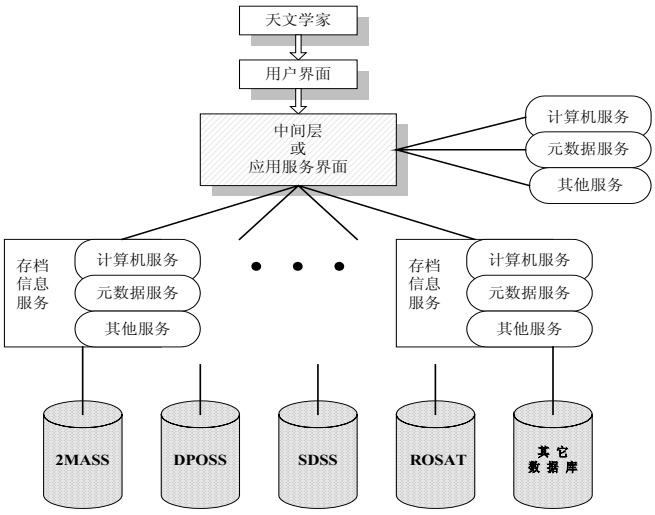


图 1.2　虚拟天文台结构示意图

结束语

综上所述，虚拟天文台是完全建筑于天文数据库和网络信息技术之上，需要强大的计算机与网络的软硬件支持，以及开发新的数据分析技术和知识发现工具。要想使虚拟天文台立于不败之地，就必须在原有数据基础上吸收新的巡天资料和数据。这样，采用库文件转换数据和转换子库之间的元数据与元服务是必要的，这要求各分析工具能协调处理来自不同数据库的数据。所有的这些工作将在

创建国家虚拟天文台或全球虚拟天文台后得以完成。虚拟天文台将把空间与地面观测设备得到的多波段巡天的海量数据有机地联合起来，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。无论在科学上还是技术上，未来的虚拟天文台将为天文学带来一场新的革命。因此，可以说虚拟天文台是推动新科学的推进器，促进新发现的催化剂。

随着天文观测技术的飞速发展和天文数据处理技术的不断提高，虚拟天文台的出现犹如雪中送炭，这对没有财力建造大型设备的发展中国家来说具有更为深刻的意义。虚拟天文台的出现将为天文学家、计算机科学家、数学家和统计学家的精诚合作提供良好的机遇，开创一条系统地完整地探索宇宙的大道，并向我们展示未来天文的前景和面貌，使天文学家不用望远镜、足不出户就可以探索宇宙的美好愿望得以实现。拥有虚拟天文台，任何地方的科学家和学生将会如虎添翼地做出重要的开创性的工作。同时，虚拟天文台的出现也向我们提出了严峻的考验，巡天技术的发展推动数据量的飞速增长，这就需要新的数据存储方法和新的分析工具以及强大的硬件支持，并需要培养下一代科学家以适应时代发展的需要。

参 考 文 献

[1] Djorgovski S G, Brunner R J, astro-ph/0006043, 2000, in press

[2] Szalay A S, Brunner R J, astro-ph/9812335

[3] Djorgovski S G, Mahabal A A, Brunner R J, *et al*. astro-ph/0012453, 2000, in press.

[4] Gilmore G, astro-ph/0011464, 2000, in press.

[5] Djorgovski S G, Brunner R J, Mahabal A A, *et al*. astro-ph/0012489, 2000, in press.

[6] Zwicky F, Morphological Astronomy,Berlin: Springer Verlag, 1957

[7] Harwit M, QJRAS, 1975, 16: 378

[8] Harwit M, Hildebrand R, Nature, 1986, 320: 724

[9] Warren S, Hewitt P, Irwin M, *et al*. Nature, 1987, 325: 131

[10] Irwin M, McMahon R, Hazard C, in The Space Distribution of Quasars, ed. D Crampton, ASPCS, 1991, 21: 117

[11] Fan X, *et al*. Astron. J., 1999, 118: 1

[12] Fan X, *et al*. Astron. J., 2000, 119: 1

[13] Fan X, *et al*. Astron. J., 2000, 120: 1167

[14] Steidel C, Adelberger K, Giavalisco M, *et al*. Ap. J., 1999, 519:

[15] Dickinson M, *et al*. Ap. J., 2000, 531: 624

[16] Kirkpatrick D, *et al*. Ap. J., 1999, 519: 802

[17] Strauss M, *et al*. Ap. J., 522: L61

[18] Burgasser A, *et al*. Ap. J., 2000, 120: 1100

[19] Fan X, *et al*. Astron. J., 2000, 119: 928

[20] Leggett S, *et al*. Ap. J., 2000, 536: L35

[21] Djorgovski S G, Gal R R, Odewahn S C, *et al*. In Wide Field Surveys in Cosmology, eds. S Colombi *et al*. Gif sur Yvette:Eds. Frontieres, 1998

[22] Wolf C, *et al*. Astron. Astrophys, 1999, 343: 399

[23] Warren S, Hewitt P, Foltz C, M.N.R.A.S, 2000,312: 827

[24] Yanny B, *et al*. Ap. J., 2000, 540: 825

[25] Odewahn S C, Windhorst R, Driver S, *et al*. Ap. J., 1996, 472: L13

[26] Brunner R G, Djorgovski S G, Gal A A, *et al*. astro-ph/0010619, 2000, in press.

[27] Brunner R G, astro-ph/0012361, 2000, in press.

[28] Schombert J, astro-ph/0009080, 2000, in press

[29] Fayyad U, Piatetsky-Shapiro G, Smyth P, *et al*. (eds.) Advances in Knowledge Discovery and Data Mining, Boston: AAAI/MIT Press

[30] Weir N, Fayyad U, Djorgovski S G, *et al*. The SKICAT System for Processing and Analysing Digital Imaging Sky Surveys, Publ. Astron. Soc. Pacific, 1995, 107: 1243

[31] Weir N, Fayyad U, Djorgovski S G, Automated Star/Galaxy Classification for Digitized POSS-II, Astron. J. 1995, 109: 2401

[32] Fayyad U, Smyth P, Weir N, *et al*. Automated Analysis and Exploration of Image Databases: Results, Progress, and Challenges, J. Intel. Inf. Sys. 1995, 4: 7

[33] Djorgovski S G, Carvalho R R, Odewahn S C, *et al*. astro-ph/9708218

[34] Goebel J, Volk K, Walker H, *et al*. AA, 1989, 222: L5

[35] De Carvalho R, Djorgovski S G, Weir N, *et al*. in Astronomical Data Analysis Software and Systems IV, eds. R Shaw *et al*. ASP Conference Series, 1995, 77: 272

[36] Yoo J, Gray A, Roden J, *et al*. in Astronomical Data Analysis Software and Systems V, eds. G Jacoby, Barnes, ASP Conference Series, 1996, 101: 41

[37] Brunner R J, Prince T, Good J, *et al*. astro-ph/0011222, 2000, in press

[38] Adanti S, Battinelli P, Capuzzo-Dolcetta R, *et al*. 1994, Astron. Astrophys. Suppl. Ser. 108: 395

[39] Connolly A J, Szalay A S, Bershady M A, *et al*. Astron. J. 1995, 110(3): 1071

[40] Connolly A J, Szalay A S, Astron. J. 1999, 117: 2052

§1.3 数据挖掘和知识发现

§1.3.1 数据挖掘和知识发现

数据挖掘技术概述

随着数据库技术的迅速发展以及数据库管理系统的广泛应用，人们积累的数据越来越多。激增的数据背后隐藏着许多重要的信息，人们希望能够对其进行更高层次的分析，以便更好地利用这些数据。目前的数据库系统可以高效地实现数据的录入、查询、统计等功能，但无法发现数据中存在的关系和规则，无法根据现有的数据预测未来的发展趋势。缺乏挖掘数据背后隐藏的知识的手段，导致了“数据爆炸但知识贫乏”的现象。数据挖掘技术是人们长期对数据库技术进行研究和开发的结果。起初各种商业数据是存储在计算机的数据库中的，然后发展到可对数据库进行查询和访问，进而发展到对数据库的即时遍历。数据挖掘使数据库技术进入了一个更高级的阶段，它不仅能对过去的数据进行查询和遍历，并且能够找出过去数据之间的潜在联系，从而促进信息的传递。

研究数据挖掘的历史，可以发现数据挖掘的快速增长是和商业数据库的空前速度增长分不开的，并且九十年代较为成熟的数据仓库正同样广泛地应用于各种商业领域。从商业数据到商业信息的进化过程中，每一步前进都是建立在上一步的基础上的。表 1.1 给出了数据进化的四个阶段，从中可以看到，第四步进化是革命性的，因为从用户的角度来看，这一阶段的数据库技术已经可以快速地回答商业上的很多问题了。

表 1.1 数据进化的四个阶段

进化阶段	时间段	技术支持	生产厂家	产品特点
数据搜集	60 年代	计算机、磁带等	IBM,CDC	提供静态历史数据
数据访问	80 年代	关系数据库、结构化查询语言 SQL	Oracle 、 Sybase 、 Informix 、IBM、Microsoft	在纪录中存在动态历史数据信息
数据仓库	90 年代	联机分析处理、多维数据库	Pilot 、 Comshare 、 Arbor 、Cognos、Microstrategy	在各层次提供回溯的动态的历史数据
数据挖掘	正在流行	高级算法、多处理系统、海量算法	Pilot、Lockheed、IBM、SGI、其他初创公司	可提供预测性信息

随着数据库技术的成熟和数据应用的普及，人类积累的数据量正在以指数速度迅速增长。进入九十年代，伴随着因特网（Internet）的出现和发展，以及随之而来的企业内部网（Intranet）和企业外部网（Extranet）以及虚拟私有网（VPN，

Virtual Private Network）的产生和应用，将整个世界联成一个小小的地球村，人们可以跨越时空地在网上交换数据信息和协同工作。这样，展现在人们面前的已不是局限于本部门、本单位和本行业的庞大数据库，而是浩瀚无垠的信息海洋，数据洪水正向人们滚滚涌来。当数据量极度增长时，如果没有有效的方法，由计算机及信息技术来提取有用的信息和知识，人们会感到面对信息海洋像大海捞针一样束手无策。据估计，一个大型企业数据库中数据，只有百分之七得到很好应用。这样，相对于“数据过剩”和“信息爆炸”，人们又感到“信息贫乏”（Information poor）和“数据关在牢笼中”(data in jail)，奈斯伯特(John Naisbett)惊呼“We are drowning in information，but starving for knowledge”（人类正被数据淹没，却饥渴于知识）。面临浩渺无际的数据，人们呼唤从数据汪洋中来一个去粗存精、去伪存真的技术。从数据库中发现知识(KDD)及其核心技术——数据挖掘(DM)便应运而生了，并得以蓬勃发展，越来越显示出其强大的生命力。

所谓数据是指有关事实的集合，记录与事物有关的原始信息。模式是一个用语言来表示的一个表达式，它可用来描述数据集的某个子集。我们所说的知识，是对数据包涵信息的更抽象的描述。对大量数据进行分析的过程，包括数据准备、模式搜索、知识评价，以及反复的修改求精；该过程要求是非平凡的，意思是要有一定程度的智能性、自动性(仅仅给出所有数据的总和不能算作是一个发现过程)。有效性是指发现的模式对于新的数据仍保持有一定的可信度。新颖性要求发现的模式应该是新的。潜在有用性是指发现的知识将来有实际效用，如用于决策支持系统可提高经济效益。最终可理解性要求发现的模式能被用户理解，目前它主要是体现在简洁性上。有效性、新颖性、潜在有用性和最终可理解性综合在一起可称之为兴趣性。

数据挖掘(Data Mining, DM)就是从大量的、不完全的、有噪声的、模糊的、随机的数据中，提取隐含在其中的、人们事先不知道的、但又是潜在有用的信息和知识的过程。确切地说，数据挖掘(DM)，又称数据库中的知识发现(Knowledge Discovery in Database, KDD)，是指从大型数据库或数据仓库中提取隐含的、未知的、非平凡的及有潜在应用价值的信息或模式，它是数据库研究中的一个很有应用价值的新领域，融合了数据库、人工智能、机器学习、统计学等多个领域的理论和技术。数据挖掘其实是知识发现的核心部分，而知识发现是在积累了大量数据后，从中识别出有效的、新颖的、潜在的、有用的及最终可以理解的知识，人们利用这些知识改进工作，提高效率和效益。数据挖掘是信息发展到一定程度的必然产物，是利用积累数据的一个高级阶段。用数据库管理系统来存储数据，用机器学习的方法来分析数据，挖掘大量数据背后的知识，这两者的结合促成了数据库中的知识发现（KDD）的产生。

由于数据挖掘是一门新兴学科，况且它又是一门受到来自各种不同领域的研究者关注的边缘学科，因此产生了很多不同的术语，除了称为“知识挖掘”外，主要还有如下若干种称法：“数据发现”、“数据开采”、“知识抽取”、“信息发现”、“知识发现”、“智能数据分析”、“探索式数据分析”、“信息收获”和“数据考古”等等。还有很多和这些术语相近似的术语，如从数据库中发现知识(KDD)、数据分析、数据融合(Data Fusion)以及决策支持等。人们把原始数据看作是形成知识的源泉，就像从矿石中采矿一样。原始数据可以是结构化的，如关系数据库中的数据，也可以是半结构化的，如文本、图形、图像数据，甚至是分布在网络上的异构型数据。发现知识的方法可以是数学的，也可以是非数学的；可以是演绎的，也可以是归纳的。发现的知识可以被用于信息管理、查询优化、决策支持、过程控制等，还可以用于数据自身的维护。因此，数据挖掘是一门很广义的交叉学科，它汇聚了不同领域的研究者，尤其是数据库、人工智能、数理统计、可视化、并行计算等方面的学者和工程技术人员。“数据挖掘”被许多研究者看作仅是知识发现的一个步骤。相对来讲，数据挖掘主要流行于统计界、数据分析、数据库和管理信息系统(MIS)界；而知识发现则主要流行于人工智能和机器学习界。

数据挖掘和知识发现虽然只有十年的历史，但它已被越来越多的领域所采用，并取得了较好效果。这些领域有科学研究、市场营销、金融投资、欺诈甄别、产品制造、通信网络管理等。由加州理工学院喷气推进实验室与天文科学家合作开发的 SKICAT(Sky Image Cataloging and Analysis Tool)是第一个获得相当成功的数据挖掘应用，已经帮助科学家发现了 16 颗极其遥远的类星体。

数据挖掘的目的是从大量数据中寻找有用的信息，它起先主要应用于商业活动，例如市场管理、风险管理和欺诈管理。它能否应用于对科学数据的加工，并从已有的科学数据库中寻找出新的科学知识或规律，是本文提出的并想探讨的问题。想法是，既然可从大量的商业活动所积累的数据中挖掘出有用的信息，那么就应该有可能从大量科研活动所积累的数据中挖掘出我们还未掌握的知识，即新的科学发现。我们预测：数据挖掘技术应该成为对科学数据加工的一种新的技术，至少应该运用这种技术对大量科学数据加工做出尝试，因此科学工作者应了解数据挖掘的技术、方法、过程和步骤，并探索其对科学数据挖掘的潜在应用或应用领域。

特别要指出的是，数据挖掘技术从一开始就是面向应用的。它不仅是面向特定数据库的简单检索查询调用，而且要对这些数据进行微观、中观乃至宏观的统计、分析、综合和推理，以指导实际问题的求解，企图发现事件间的相互关联，甚至利用已有的数据对未来的活动进行预测。例如加拿大 BC 省电话公司要求加拿大 SimonFraser 大学 KDD 研究组，根据其拥有十多年的客户数据，总结、分

析并提出新的电话收费和管理办法，制定既有利于公司又有利于客户的优惠政策。这样一来，就把人们对数据的应用，从低层次的末端查询操作，提高到为各级经营决策者提供决策支持。这种需求驱动力，比数据库查询更为强大。同时需要指出的是，这里所说的知识发现，不是要求发现放之四海而皆准的真理，也不是要去发现崭新的自然科学定理和纯数学公式，更不是什么机器定理证明。所有发现的知识都是相对的，是有特定前提和约束条件、面向特定领域的，同时还要能够易于被用户理解，最好能用自然语言表达发现结果。因此 DM/KDD 的研究成果是很讲求实际的。

虽然数据挖掘和知识发现已经受到许多关注并取得了广泛应用，但它仍处于发展的早期，还有很多研究难题和面临的挑战，如数据的巨量性、动态性、噪声性、缺值和稀疏性，发现模式的可理解性、兴趣性或价值性，应用系统的集成，用户的交互操作，知识的更新管理，复杂数据库的处理等等。

数据挖掘的商业背景

数据挖掘首先是需要商业环境中收集了大量的数据，然后要求挖掘的知识是有价值的。有价值对商业而言，不外乎三种情况：降低开销；提高收入；增加股票价格。具体应用：

⑴ 数据挖掘作为研究工具（Research）；

⑵ 数据挖掘提高过程控制（Process Improvement）；

⑶ 数据挖掘作为市场营销工具（Marketing）；

⑷ 数据挖掘作为客户关系管理 CRM 工具(Customer Relationship Management，CRM)。

数据挖掘的技术背景

⑴ 数据挖掘技术包括三个主要部分：算法和技术、数据、建模能力。

⑵ 数据挖掘与机器学习（Machine Learning）：机器学习是计算机科学和人工智能 AI 发展的产物。机器学习分为两种学习方式：自组织学习（如神经网络）；从例子中归纳出规则（如决策树）。数据挖掘是八十年代，投资 AI 研究项目失败后，AI 转入实际应用时提出的。它是一个新兴的，面向商业应用的 AI 研究。选择数据挖掘这一术语，表明了与统计、精算、长期从事预言模型的经济学家之间没有技术的重叠。

⑶ 数据挖掘与统计：统计也开始支持数据挖掘。统计本包括预言算法（回归）、抽样、基于经验的设计等。

⑷ 数据挖掘与决策支持系统：数据仓库、联机分析处理（OLAP）、数据集

市（Data Mart）、多维数据库、决策支持工具融合。将数据仓库、OLAP、数据挖掘融合在一起，构成企业决策分析环境。

数据挖掘的社会背景

数据挖掘与个人预言：数据挖掘号称能通过历史数据的分析，预测客户的行为，而事实上，客户自己可能都不明确自己下一步要做什么。所以，数据挖掘的结果，没有人们想象中那样神秘，它不可能是完全正确的。客户的行为是与社会环境相关连的，所以数据挖掘本身也受社会背景的影响。比如说，在美国对银行信用卡客户信用评级的模型运行得非常成功，但是，它可能不适合中国。

数据挖掘和知识发现研究的兴起

近年来已有越来越多的研究者加入到数据库知识发现 KDD 研究的行列。KDD 也称数据挖掘(DM), 它的研究引起了人们的极大兴趣。自 1989 年始已经举行了四届有关 KDD 专题的国际讨论会,并且取得了一些相当有意义的成果。

随着大量的大规模的数据库迅速不断地增长, 人们对数据库的应用已不满足于仅对数据库进行查询和检索。仅用查询检索不能帮助用户从数据中提取带有结论性的有用信息。这样数据库中蕴藏的丰富知识, 就得不到充分的发掘和利用。从而造成了信息的浪费, 由此也会产生大量的数据垃圾。

从人工智能应用来看, 专家系统的研究虽然取得了一定的进展。但是, 知识获取仍然是专家系统研究中的瓶颈。知识工程师从领域专家处获取知识是非常复杂的个人到个人之间的交互过程，具有很强的个性, 没有统一的办法。因此, 有必要考虑从数据库中自动挖掘新的知识。这些都需要新的数据处理技术，KDD 便应运而生。KDD 的研究内容：是能自动地处理数据库中大量的原始数据, 从中挖掘搜索出具有必然性的、富有意义的模式(pattern)。KDD 的一个主要问题是数据库中潜在的可能关系模式的数量太大，因此要想搜索到有用的模式，必须借用人工智能技术，特别是来自机器学习领域的方法。

科技信息是巨大的社会财富，科学数据库是将科技信息转化为生产力的重要手段。由于各种科技数据和文献的急剧增长，利用传统的工具和方法已不能有效地处理和传播，于是不得不求助于计算机和先进的通信技术，从而大大推进了科学数据库及信息技术的发展和应用。随着科学数据库的发展，数据库知识发现必然会发挥极其重要的作用，从各种学科数据库获取有用的知识，发现各门学科数据所反映的规律性。

数据挖掘和知识发现的研究现状

KDD 一词首次出现在 1989 年 8 月举行的第 11 届国际联合人工智能学术会议上。迄今为止，由美国人工智能协会主办的 KDD 国际研讨会已经召开了 7 次，规模由原来的专题讨论会发展到国际学术大会，人数由二三十人到七八百人，论文收录比例从 2:1 到 6:1，研究重点也逐渐从发现方法转向系统应用，并且注重多种发现策略和技术的集成，以及多种学科之间的相互渗透。其他内容的专题会议也把数据挖掘和知识发现列为议题之一，成为当前计算机科学界的一大热点。

此外，数据库、人工智能、信息处理、知识工程等领域的国际学术刊物也纷纷开辟了 KDD 专题或专刊。IEEE 的 Knowledge and Data Engineering 会刊领先在 1993 年出版了 KDD 技术专刊，所发表的 5 篇论文代表了当时 KDD 研究的最新成果和动态，较全面地论述了 KDD 系统方法论、发现结果的评价、KDD 系统设计的逻辑方法，集中讨论了鉴于数据库的动态性冗余、高噪声和不确定的 KDD 系统与其它传统的机器学习、专家系统、人工神经网络、数理统计分析系统的联系和区别，以及相应的基本对策。6 篇论文摘要展示了 KDD 在从建立分子模型到设计制造业的具体应用。

不仅如此，在 Internet 上还有不少 KDD 电子出版物，其中以半月刊 Knowledge Discovery Nuggets 最为权威，如要免费订阅，只需向 http://www.kdnuggets.com/subscribe.html 发送一份电子邮件即可，还可以下载各种各样的数据挖掘工具软件和典型的样本数据仓库，供人们测试和评价。另一份在线周刊为 DS*(DS 代表决策支持)，1997 年 10 月 7 日开始出版，可向 dstrial@tgc.com 提出免费订阅申请。在网上，还有一个自由论坛 DMEmailClub，人们通过电子邮件相互讨论 DM/KDD 的热点问题。而领导整个潮流的 DM/KDD 开发和研究中心，当数设在美国 EMDEN 的 IBM 公司开发部。

随着 DM/KDD 研究逐步走向深入，人们越来越清楚地认识到，DM/KDD 的研究主要有 3 个技术支柱，即数据库、人工智能和数理统计。

数据库技术在经过了 80 年代的辉煌之后，已经在各行各业成为一种数据库文化或时尚，数据库界目前除了关注分布式数据库、面向对象数据库、WEB 数据库、多媒体数据库、查询优化和并行计算等技术外，已经开始反思数据库最实质的应用仅仅是查询吗？理论根基最深的关系数据库最本质的技术进步，就是数据存放和数据使用之间的相互分离。查询是数据库的奴隶，发现才是数据库的主人；数据只为职员服务，不为老板服务！这是很多单位的领导在热心数据库建设后发出的感叹。

由于数据库文化的迅速普及，用数据库作为知识源具有坚实的基础；另一方面，对于一个感兴趣的特定领域——客观世界，先用数据库技术将其形式化并组织起来，就会大大提高知识获取起点，以后从中发掘或发现的所有知识都是针对该数据库而言的。因此，在需求的驱动下，很多数据库学者转向对数据仓库和数据挖掘的研究，从对演绎数据库的研究转向对归纳数据库的研究。

专家系统曾经是人工智能研究工作者的骄傲。专家系统实质上是一个问题求解系统，目前的主要理论工具是基于谓词演算的机器定理证明技术——二阶演绎系统。领域专家长期以来面向一个特定领域的经验世界，通过人脑的思维活动积累了大量的有用信息。

在研制一个专家系统时，知识工程师首先要从领域专家那里获取知识，这一过程实质上是归纳过程，是非常复杂的个人到个人之间的交互过程，有很强的个性和随机性。因此，知识获取成为专家系统研究中公认的瓶颈问题。

其次，知识工程师在整理表达从领域专家那里获得的知识时，用 if-then 等类的规则来表达，约束性太大；用常规数理逻辑来表达社会现象和人的思维活动局限性太大，也太困难；勉强抽象出来的规则有很强的工艺色彩，差异性极大，知识表示又成为一大难题。此外，即使某个领域的知识通过一定手段获取并表达了，但这样做成的专家系统对常识和百科知识出奇地贫乏，而人类专家的知识是以拥有大量常识为基础的。人工智能学家 Feigenbaum 估计，一般人拥有的常识存入计算机大约有 100 万条事实和抽象经验法则，离开常识的专家系统有时会比傻子还傻。例如战场指挥员会根据“在某地发现一只刚死的波斯猫”的情报很快断定敌高级指挥所的位置，而再好的军事专家系统也难以顾全到如此的信息。

以上这 3 大难题大大限制了专家系统的应用，使得专家系统目前还停留在构造诸如发动机故障论断一类的水平上。人工智能学者开始着手基于案例的推理，尤其是从事机器学习的科学家们，不再满足自己构造的小样本学习模式的象牙塔，开始正视现实生活中大量的、不完全的、有噪声的、模糊的、随机的大数据样本，也走上了数据挖掘的道路。数理统计是应用数学中最重要、最活跃的学科之一，它在计算机发明之前就诞生了，迄今已有几百年的发展历史。如今相当强大有效的数理统计方法和工具，已成为信息咨询业的基础。信息时代，咨询业更为发达。然而，数理统计和数据库技术结合得并不算快，数据库查询语言 SQL 中的聚合函数功能极其简单，就是一个证明。咨询业用数据库查询数据还远远不够。一旦人们有了从数据查询到知识发现、从数据演绎到数据归纳的要求，概率论和数理统计就获得了新的生命力，所以才会在 DM/KDD 这个结合点上，立即呈现出“忽如一夜春风来，千树万树梨花开”的繁荣景象。

什么是数据挖掘？

数据挖掘的定义非常模糊，对它的定义取决于定义者的观点和背景。如下是一些 DM 文献中的定义：

数据挖掘是一个从大型数据库中提取以前未知的、可理解的、可执行的信息并用它来进行关键的商业决策的过程。（Zekulin）

数据挖掘是用在知识发现过程中来辩识存在于数据中的未知关系和模式的一些方法。（Ferruzza）

数据挖掘是发现数据中有益模式的过程。（John）

数据挖掘是我们为那些未知的信息模式而研究大型数据集的一个决策支持过程。（Parsaye）

数据挖掘是决策树、神经网络、规则推断、最近邻方法、遗传算法。（Mehta）

数据挖掘的定义几经变动，最新的描述性定义是由 Usama M. Fayyad 等[1]给出的：数据挖掘是从数据集中识别出有效的、新颖的、潜在有用的，以及最终可理解的模式的非平凡过程。

通常认为数据挖掘（DM）就是从大量的、不完全的、有噪声的、模糊的、随机的实际应用数据中，提取隐含在其中的、人们事先不知道的、但又是潜在有用的信息和知识的过程。

数据挖掘技术的演变

数据挖掘其实是一个逐渐演变的过程, 电子数据处理的初期, 人们就试图通过某些方法来实现自动决策支持, 当时机器学习成为人们关心的焦点。机器学习的过程就是将一些已知的并已被成功解决的问题作为范例输入计算机, 机器通过学习这些范例总结并生成相应的规则, 这些规则具有通用性, 使用它们可以解决某一类的问题。随着神经网络技术的形成和发展, 人们的注意力转向知识工程, 知识工程不同于机器学习那样给计算机输入范例，让它生成规则，而是直接给计算机输入已被代码化的规则，而计算机是通过使用这些规则来解决某些问题。专家系统就是这种方法所得到的成果，但它存在投资大、效果不甚理想等不足。80 年代人们又在新的神经网络理论的指导下，重新回到机器学习的方法上，并将其成果应用于处理大型商业数据库。随着在 80 年代末一个新的术语，它就是数据库中的知识发现（KDD）。它泛指所有从源数据中发掘模式或联系的方法，人们接受了这个术语，并用 KDD 来描述整个数据挖掘的过程，包括最开始的制定业务目标到最终的结果分析，而用数据挖掘来描述使用挖掘算法进行数据挖掘的子过程。但最近人们却逐渐开始发现数据挖掘中有许多工作可以由统计方法来完成，并认为最好的策略是将统计方法与数据挖掘有机地结合起来。

数据仓库技术的发展与数据挖掘有着密切的关系。数据仓库的发展是促进数据挖掘越来越热的原因之一。但是，数据仓库并不是数据挖掘的先决条件，因为有很多数据挖掘可直接从操作数据源中挖掘信息。

知识发现的核心—数据挖掘

在数据库领域习惯叫“数据库中的知识发现”（KDD），它强调“知识”是一个由数据导出发现的最终结果。KDD 实际上是智能技术与数据库技术的结合。KDD 也就是从数据库中提取有价值知识的过程，是数据库技术与机器学习学科的交叉。数据库技术侧重于对数据存储处理的高效率方法的研究，而机器学习技术则侧重于设计新的方法从数据中提取知识。KDD 利用数据库技术对数据进行前端处理，并利用机器学习方法从处理后的数据中提取有用的知识。

KDD 是一门交叉学科，涉及到人工智能、机器学习、模式识别、统计学、智能数据库、知识获取、数据可视化、专家系统等多个领域。KDD 的发展领域并不单一：学习算法效率和可扩充性是其发展极为重要的两个方面；由于 KDD 所处理的数据来自现实世界，数据的完整性、一致性和正确性都很难得到保证，如何将这些数据加工成学习算法可以接收的数据，是一个值得深入研究的方向；利用目前数据库技术所取得的研究成果来加快学习过程、提高学习效率，又是一个课题；有效结合与数据库数据有关的领域知识和专家学者的经验，是 KDD 提高学习算法效率的一个关键。事实上，KDD 的发展空间很大，从某种意义上说，KDD 才刚刚起步。

从定义中可以看出，KDD 是一个高级的处理过程，它从数据集中识别出以模式来表示的知识。高级的处理过程是指一个多步骤的处理过程，多步骤之间相互影响、反复调整，形成一种螺旋式的上升过程。数据挖掘是 KDD 的最核心部分，是采用机器学习、统计等方法进行知识学习的阶段。数据挖掘与传统分析工具不同点在于数据挖掘使用的是基于发现的方法，运用模式匹配和其它算法决定数据之间的重要联系。数据挖掘的任务是从数据中发现模式。数据挖掘算法的好坏将直接影响到所发现知识的好坏。目前大多数的研究都集中在数据挖掘算法和应用上。需要说明的是，有的学者认为，数据挖掘和知识发现含义相同，表示成 DM/KDD。它是一个反复的过程，通常包含多个相互联系的步骤：数据准备、数据选择、数据预处理、数据缩减、KDD 目标确定、选取算法、提取规则、数据挖掘、模式解释及知识评价等，从而得到知识，最后应用。该模型强调了 KDD 需要领域专家的参与，由专业知识指导数据库中的知识发现的各个阶段，并对发现知识进行评价。有的模型是以用户为中心的，这样的模型更着重于对用户进行知识发现的整个过程的支持，而不是仅仅限于在数据挖掘的一个阶段上，交互市

场分析及分类系统就是这样的例子，该系统特别强调对用户与数据库交互的支持。此外，还有的 KDD 模型是交互式的迭代过程，它包含许多由用户做决策的步骤，这样的模型强调处理过程的反复性，如再求精分类、粗糙集和模糊集抽取规则，神经网络的反复训练等。数据库的知识发现的交叉性、综合性，可以从它所利用的技术来体现，这些技术主要是分类技术、聚类技术、神经网络技术、粗糙集技术、统计技术和关联规则技术等。在实际中，人们往往不严格区分数据挖掘和数据库中的知识发现，把两者混淆使用。一般在科研领域中称为知识发现（KDD），而在工程领域则称为数据挖掘（DM）。

数据挖掘和知识发现过程

数据挖掘是指一个完整的过程，该过程从大型数据库中挖掘先前未知的、有效的、可实用的信息，并使用这些信息做出决策或丰富知识。

数据挖掘与传统的数据分析(如查询、报表、联机应用分析)的本质区别是数据挖掘是在没有明确假设的前提下去挖掘信息、发现知识。数据挖掘所得到的信息应具有先前未知、有效和可实用三个特征。

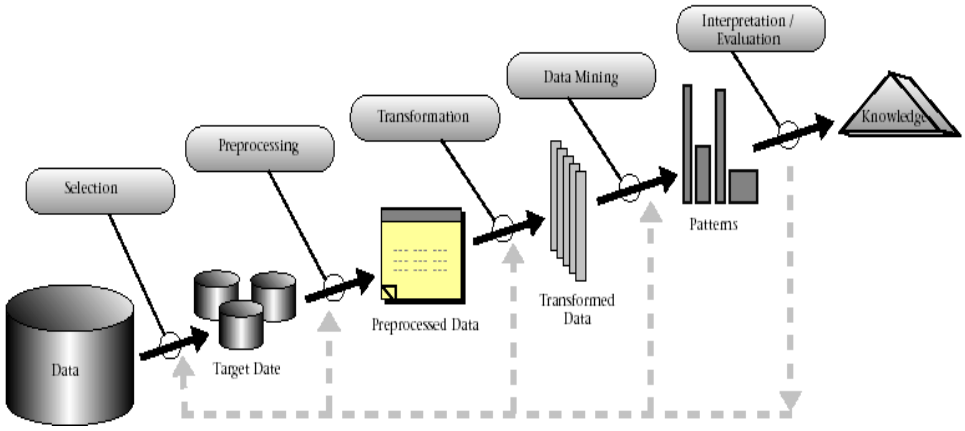
先前未知的信息是指该信息是预先未曾预料到的，即数据挖掘是要发现那些不能靠直觉发现的信息或知识，甚至是违背直觉的信息或知识，挖掘出的信息越是出乎意料，就可能越有价值。在商业应用中最典型的例子就是一家连锁店通过数据挖掘发现了小孩尿布和啤酒之间有着惊人的联系。

信息的有效性要求挖掘前要对被挖掘的数据进行仔细检查，保证它们的有效性，才能保证挖掘出来的信息的有效性。从某种程度来讲，科学数据的有效性与其它数据相比往往是能得到保证的。

最为重要的是要求所得的信息具有可实用性，即这些信息或知识对于所讨论的业务或研究领域是有效的、有实用价值和可实现的。常识性的结论，或也被人们或竞争对手早已掌握的或无法实现的事实都是没有意义的。

数据挖掘和知识发现（DM/KDD）是从数据中发现有用知识的整个过程；数据挖掘（DM）是 KDD 过程中的一个特定步骤，它用专门算法从数据中抽取模式（patterns）发现知识。

KDD 过程是多个步骤相互连接、反复进行人机交互的过程。知识发现的基本过程包括确定业务对象；数据准备；数据挖掘；结果解释和评估；知识的同化。图 1.3 描述了知识发现的基本过程和主要步骤[2]。



数据选择　数据预处理　数据转换　数据挖掘　解释和评估

图 1.3　知识发现过程的步骤

数据挖掘过程中各步骤的大体内容如下：

⑴ 确定挖掘对象

清晰地定义出问题, 认清数据挖掘的目的是数据挖掘的重要一步。挖掘的最后结构是不可预测的, 但要探索的问题应是有预见的, 为了数据挖掘而数据挖掘则带有盲目性, 是不会成功的。

⑵数据准备

① 数据的选择

搜索所有与挖掘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数据集的子集上聚焦，挑出适用于数据挖掘应用的数据。

② 数据的预处理

去除噪声或无关数据，去除空白数据域，考虑时间顺序和数据变化等。研究数据的质量,为进一步的分析做准备，并确定将要进行的挖掘操作的类型。

③ 数据的转换

找到数据的特征表示，用维变换或转换方法减少有效变量的数目或找到数据的不变式。将数据转换成一个分析模型，这个分析模型是针对挖掘算法建立的。建立一个真正适合挖掘算法的分析模型是数据挖掘成功的关键。

⑶ 数据挖掘

对所得到的经过转换的数据进行挖掘。用 KDD 过程中的准则，选择某个特定数据挖掘算法（如汇总、分类、回归、聚类等）用于搜索数据中的模式。除了完善从选择合适的挖掘算法外，其余一切工作都能自动地完成。然后搜索或产生一个特定的感兴趣的模式或一个特定的数据集。

⑷ 结果解释和评估

解释并评估结果。解释某个发现的模式，去掉多余的不切题意的模式，转换

某个有用的模式，以使用户明白。其使用的分析方法一般应由数据挖掘操作而定，通常会用到可视化技术。

⑸ 知识的同化

将分析所得到的知识集成到业务信息系统的组织结构中去，获得这些知识的作用或证明这些知识。用预先、可信的知识检查和解决知识中可能的矛盾。

数据挖掘过程的分步实现，不同的步会需要不同专长的人员，他们大体可以分为三类。

分析人员：要求精通业务，能够解释业务对象，并根据各挖掘对象确定出用于数据定义和挖掘算法的挖掘需求。

数据分析人员：精通数据分析技术，并对统计学有较熟练的掌握，有能力把挖掘需求转化为数据挖掘的各步操作，并为每步操作选择合适的技术。

数据管理人员：精通数据管理技术，并从数据库或数据仓库中收集数据。

从上可见，数据挖掘是一个多种专家合作的过程，也是一个在资金上和技术上高投入的过程。

在数据挖掘中被研究的业务对象是整个过程的基础，它驱动了整个数据挖掘过程，也是检验最后结果和指引分析人员完成数据挖掘的依据和顾问。图 1.3 各步骤是按一定顺序完成的，当然整个过程中还会存在步骤间的反馈。数据挖掘的过程并不是自动的，绝大多数的工作需要人工完成。图 1.4 给出了各步骤在整个过程中的工作量之比。可以看到，60%的时间用在数据准备上，这说明了数据挖掘对数据的严格要求，而后挖掘工作仅占总工作量的 10%。

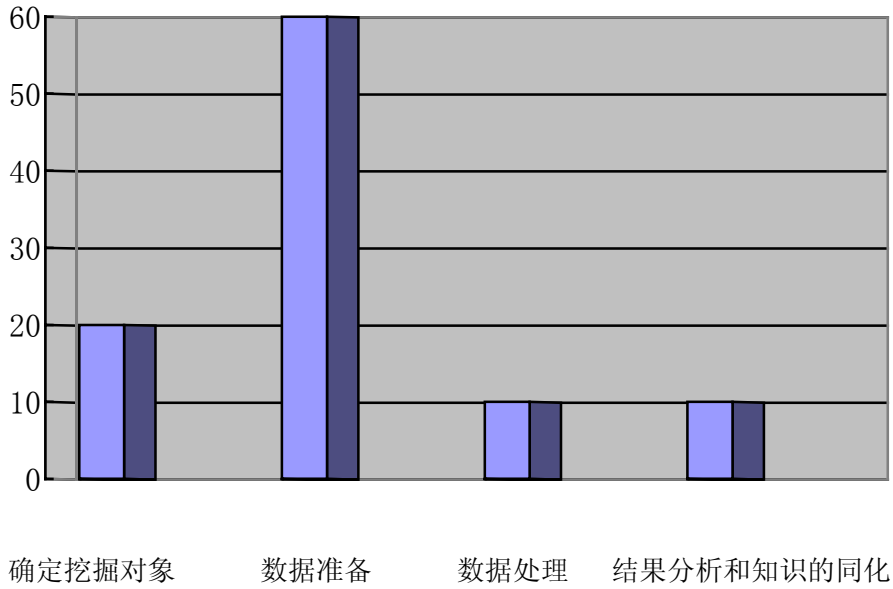


图 1.4　数据挖掘过程工作量比例

数据挖掘的任务和方法

数据挖掘的核心模块技术历经了数十年的发展，其中包括数理统计、人工智能、机器学习。今天，这些成熟的技术，加上高性能的关系数据库引擎以及广泛的数据集成，让数据挖掘技术在当前的数据仓库环境中进入了实用阶段。数据挖掘中要分析的数据的范围是非常广泛的,从自然科学、社会科学、商业数据，到科学处理产生的数据或卫星观测得到的数据。它们的数据表示也各种各样，有关系型，也有层次型。由于关系数据库应用广,具有规整统一的组织结构,通用的查询语言,特别是关系之间及属性之间具有平等性的优点。因此，目前数据挖掘的主要对象仍是关系数据库。

数据挖掘所能发现的知识有如下几种：广义型知识，反映同类事物共同性质的知识；特征型知识，反映事物各方面特征的知识；差异型知识，反映不同事物之间属性差别的知识；关联型知识，反映事物之间依赖或关联的知识；预测型知识，根据历史的和当前的数据推测未来数据；偏离型知识，揭示事物偏离常规的异常现象。所有这些知识都可以在不同的概念层次上被发现，随着概念树的提升，从微观到中观再到宏观，以满足不同用户、不同层次决策的需要。例如，从一家超市的数据仓库中，可以发现的一条典型关联规则可能是“买面包和黄油的顾客十有八九也买牛奶”，也可能是“买食品的顾客几乎都用信用卡”，这种规则对于商家开发和实施客户化的销售计划和策略是非常有用的。数据挖掘可发现的知识也有各种表示形式,如法则(RULES)、规则(REGULARITY)、科学定律、方程或概念网等等。

数据挖掘利用的技术越多，得出的结果的精确性就越高。原因很简单，对于某一种技术不适用的问题，其它方法即可能奏效，这主要取决于问题的类型以及数据的类型和规模。知识发现过程主要有三个步骤：用户定义要发现的问题；系统根据问题进行数据搜索、模式抽取；评价所发现的知识的质量的好坏。三者之中，核心技术是第二步，即搜索及模式抽取方法。

数据挖掘涉及的学科领域和方法很多，有多种分类法。根据挖掘任务分，可分为分类或预测模型发现、数据总结、聚类、关联规则发现、序列模式发现、依赖关系或依赖模型发现、异常和趋势发现等等；根据挖掘对象分，有关系数据库、面向对象数据库、空间数据库、时态数据库、文本数据源、多媒体数据库、异质数据库、遗产数据库以及环球网 Web；根据挖掘方法分，可分为：机器学习方法、统计方法、神经网络方法和数据库方法。机器学习中，可细分为：归纳学习方法(决策树、规则归纳等)、基于范例学习、遗传算法等。统计方法中，可细分为：回归分析(多元回归、自回归等)、判别分析(贝叶斯判别、费歇尔判别、非参数判别等)、聚类分析(系统聚类、动态聚类等)、探索性分析(主分量分析法、相关分

析法等)等。神经网络方法中，可细分为：前向神经网络(BP 算法等)、自组织神经网络(自组织特征映射、竞争学习等)等。数据库方法主要是多维数据分析或联机分析处理（OLAP）方法，另外还有面向属性的归纳方法。

以下将主要从挖掘任务和挖掘方法的角度，着重讨论数据总结、分类分析、聚类分析、关联规则分析、序列模式分析、概念描述、偏差检测等七种非常重要的发现任务和挖掘方法。

⑴ 数据总结

数据总结目的是对数据进行浓缩，给出它的紧凑描述。传统的也是最简单的数据总结方法是计算出数据库的各个字段上的求和值、平均值、方差值等统计值，或者用直方图、饼状图等图形方式表示。数据挖掘主要关心从数据泛化的角度来讨论数据总结。数据泛化是一种把数据库中的有关数据从低层次抽象到高层次上的过程。由于数据库上的数据或对象所包含的信息总是最原始、基本的信息(这是为了不遗漏任何可能有用的数据信息)。人们有时希望能从较高层次的视图上处理或浏览数据，因此需要对数据进行不同层次上的泛化以适应各种查询要求。数据泛化目前主要有两种技术: 多维数据分析方法和面向属性的归纳方法。

多维数据分析方法是一种数据仓库技术，也称作联机分析处理(OLAP)。数据仓库是面向决策支持的、集成的、稳定的、不同时间的历史数据集合。决策的前提是数据分析。在数据分析中经常要用到诸如求和、总计、平均、最大、最小等汇集操作，这类操作的计算量特别大。因此一种很自然的想法是把汇集操作结果预先计算并存储起来，以便于决策支持系统使用。存储汇集操作结果的地方称作多维数据库。

采用多维数据分析方法进行数据总结，它针对的是数据仓库，数据仓库存储的是脱机的历史数据。为了处理联机数据，研究人员提出了一种面向属性的归纳方法。它的思路是直接对用户感兴趣的数据视图(用一般的标准化查询语言 SQL 即可获得)进行泛化，而不是象多维数据分析方法那样预先就存储好了泛化数据。方法的提出者对这种数据泛化技术称之为面向属性的归纳方法。原始关系经过泛化操作后得到的是一个泛化关系，它从较高的层次上总结了在低层次上的原始关系。有了泛化关系后，就可以对它进行各种深入的操作而生成满足用户需要的知识，如在泛化关系基础上生成特性规则、判别规则、分类规则，以及关联规则等。

⑵ 分类分析

设有一个数据库和一组具有不同特征的类别(标记)，该数据库中的每一个记录都赋予一个类别的标记，这样的数据库称为示例数据库或训练集。分类分析就是通过分析示例数据库中的数据，为每个类别做出准确的描述或建立分类模型或挖掘出分类规则(也常常称作分类器)，然后用这个分类规则把数据库中的数据项

映射到给定类别中的某一个，从而对数据库中的记录进行分类。分类在数据挖掘中是一项非常重要的任务，目前在商业上应用最多。分类和回归都可用于预测。预测的目的是从历史数据纪录中自动推导出对给定数据的推广描述，从而能对未来数据进行预测。与回归方法不同的是，分类的输出是离散的类别值，而回归的输出则是连续数值。这里我们将不讨论回归方法。

要构造分类器，需要有一个训练样本数据集作为输入。训练集由一组数据库记录或元组构成，每个元组是一个由有关字段(又称属性或特征)值组成的特征向量，此外，训练样本还有一个类别标记。一个具体样本的形式可为:(v1，v2，...，vn；c)；其中 vi 表示字段值，c 表示类别。

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)，对应的知识表示则为判别函数和原型事例。机器学习方法包括决策树法和规则归纳法，前者对应的表示为决策树或判别树，后者则一般为产生式规则。神经网络方法主要是 BP 算法，它的模型表示是前向反馈神经网络模型(由代表神经元的节点和代表联接权值乘积的和组成的一种体系结构)，BP 算法本质上是一种非线性判别函数。另外，最近又兴起了一种新的方法：粗糙集(rough set)，其知识表示是产生式规则。

不同的分类器有不同的特点。有三种分类器评价或比较尺度：①预测准确度；②计算复杂度；③模型描述的简洁度。预测准确度是用得最多的一种比较尺度，特别是对于预测型分类任务，目前公认的方法是 10 番分层交叉验证法。计算复杂度依赖于具体的实现细节和硬件环境，在数据挖掘中，由于操作对象是巨量的数据库，因此空间和时间的复杂度问题将是非常重要的一个环节。对于描述型的分类任务，模型描述越简洁越受欢迎，例如，采用规则表示的分类器构造法就更有用，而神经网络方法产生的结果就难以理解。

另外要注意的是，分类的效果一般与数据的特点有关，有的数据噪声大，有的有缺值，有的分布稀疏，有的字段或属性间相关性强，有的属性是离散的，而有的是连续值或混合式的。目前普遍认为不存在某种方法能适合于各种特点的数据。

⑶ 聚类分析

与分类分析不同，聚类分析输入的是一组未分类记录，并且这些记录应分成几类事先也不知道。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。它所采用的分类规则是由聚类分析工具决定的。聚类是把一组个体按照相似性归成若干类别，即“物以类聚”。它的目的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。聚类增强了人们对客观现实的认识，是概念描述

和偏差分析的先决条件。

聚类分析和分类分析是一个互逆的过程。例如在最初的分析中，分析人员根据以往的经验将要分析的数据进行标定，划分类别，然后用分类分析方法分析该数据集合，挖掘出每个类别的分类规则；接着用这些分类规则重新对这个集合(抛弃原来的划分结果)进行划分，以获得更好的分类结果。这样分析人员可以循环使用这两种分析方法直至获得满意的结果。

聚类分析的方法很多，其中包括统计方法、机器学习方法、神经网络方法、面向数据库的方法、系统聚类法、分解法、加入法、动态聚类法、模糊聚类法、运筹方法等。采用不同的聚类方法，对于相同的记录集合可能有不同的划分结果。

在统计方法中，聚类又称聚类分析，它是多元数据分析的三大方法之一(其它两种是回归分析和判别分析)。它主要研究基于几何距离的聚类，如欧式距离、明考夫斯基距离等。传统的统计聚类分析方法包括系统聚类法、分解法、加入法、动态聚类法、有序样品聚类、有重叠聚类和模糊聚类等。这种聚类方法是一种基于全局比较的聚类，它需要考察所有的个体才能决定类的划分；因此它要求所有的数据必须预先给定，而不能动态增加新的数据对象。聚类分析方法不具有线性的计算复杂度，难以适用于数据库非常大的情况。

在机器学习中聚类称作无监督或无教师归纳；因为与分类学习相比，分类学习的例子或数据对象有类别标记，而要聚类的例子则没有标记，需要由聚类学习算法来自动确定。很多人工智能文献中，聚类也称概念聚类；因为这里的距离不再是统计方法中的几何距离，而是根据概念的描述来确定的。当聚类对象可以动态增加时，概念聚类则称是概念形成。

在神经网络中，有一类无监督学习方法: 自组织神经网络方法，如 Kohonen 自组织特征映射网络、竞争学习网络等等。在数据挖掘领域里，见报道的神经网络聚类方法主要是自组织特征映射方法，IBM 在其发布的数据挖掘白皮书中就特别提到了使用此方法进行数据库聚类分割。

⑷ 关联规则分析

数据关联是数据库中存在的一类重要的可被发现的知识。若两个或多个变量的取值之间存在某种规律性，就称为关联。关联可分为简单关联、时序关联、因果关联。关联分析的目的是找出数据库中隐藏的关联网。从而为某些决策提供必要的支持。关联分析，即利用关联规则进行数据挖掘。在数据挖掘研究领域，对于关联分析的研究开展得比较深入，人们提出了多种关联规则的挖掘算法，如 APRIORI、STEM、AIS、DHP 等算法。关联分析能发现数据库中形如“90%的顾客在一次购买活动中购买商品 A 的同时购买商品 B”之类的知识。用于关联规则发现的主要对象是事务型数据库，其中针对的应用则是售货数据，也称货篮

数据。一个事务一般由如下几个部分组成: 事务处理时间、一组顾客购买的物品、有时也有顾客标识号(如信用卡号)。

由于条形码技术的发展，零售部门可以利用前端收款机收集存储大量的售货数据。因此，如果对这些历史事务数据进行分析，则可对顾客的购买行为提供极有价值的信息。例如，可以帮助如何摆放货架上的商品(如把顾客经常同时买的商品放在一起)，帮助如何规划市场(怎样相互搭配进货)。由此可见，从事务数据中发现关联规则，对于改进零售业等商业活动的决策非常重要。

设 I={i1，i2，...，im}是一组物品集(一个商场的物品可能有上万种)，D 是一组事务集(称之为事务数据库)。D 中的每个事务 T 是一组物品，显然满足 T⊆I。如果物品集 X⊆T，则称事务 T 支持物品集 X。关联规则是如下形式的一种蕴含：X→Y，其中 X⊆I，Y⊆I，且 X∩Y=I。

① 称物品集 X 具有大小为 s 的支持度，如果 D 中有 s%的事务支持物品集 X；

② 称关联规则 X→Y 在事务数据库 D 中具有大小为 s 的支持度，如果物品集 X∪Y 的支持度为 s；

③ 称规则 X→Y 在事务数据库 D 中具有大小为 c 的可信度，如果 D 中支持物品集 X 的事务中有 c%的事务同时也支持物品集 Y。

如果不考虑关联规则的支持度和可信度，那么在事务数据库中存在无穷多的关联规则。事实上，人们一般只对满足一定的支持度和可信度的关联规则感兴趣。在文献中，一般称满足一定要求(如较大的支持度和可信度)的规则为强规则。因此，为了发现有意义的关联规则，需要给定两个阈值: 最小支持度和最小可信度。前者即用户规定的关联规则必须满足的最小支持度，它表示了一组物品集在统计意义上需满足的最低程度；后者即用户规定的关联规则必须满足的最小可信度，它反应了关联规则的最低可靠度。

在实际情况下，一种更有用的关联规则是泛化关联规则。因为物品概念间存在一种层次关系，如夹克衫、滑雪衫属于外套类，外套、衬衣又属于衣服类。有了层次关系后，可以帮助发现一些更多的有意义的规则。例如：“买外套，买鞋子”(此处，外套和鞋子是较高层次上的物品或概念，因而该规则是一种泛化的关联规则)。由于商店或超市中有成千上万种物品，平均来讲，每种物品(如滑雪衫)的支持度很低，因此有时难以发现有用规则；但如果考虑到较高层次的物品(如外套)，则其支持度就较高，从而可能发现有用的规则。

另外，关联规则发现的思路还可以用于序列模式发现。用户在购买物品时，除了具有上述关联规律，还有时间上或序列上的规律，因为，很多时候顾客会这次买这些东西，下次买同上次有关的一些东西，接着又买有关的某些东西。

⑸ 序列模式分析

序列模式分析和关联分析相似，其目的也是为了挖掘数据之间的联系，但序列模式分析的侧重点在于分析数据间的前后序列关系。它能发现数据库中形如“在某一段时间内，顾客购买商品 A，接着购买商品 B，而后购买商品 C，即序列 A→B→C 出现的频度较高”之类的知识，序列模式分析描述的问题是：在给定交易序列数据库中，每个序列是按照交易时间排列的一组交易集，挖掘序列函数作用在这个交易序列数据库上，返回该数据库中出现的高频序列。在进行序列模式分析时，同样也需要由用户输入最小置信度 C 和最小支持度 S。

⑹ 依赖关系分析

数据依赖关系代表一类重要的可发现的知识。一个依赖关系存在于两个元素之间。如果一个元素 A 的值可以推出另一个元素 B 的值(A→B)，则称 B 依赖于 A。这个元素可以是字段, 也可以是字段间的关系。在发现系统中, 依赖关系分析的结果有时可以直接提供给终端用户。然而, 通常强的依赖关系反映的是固有的领域结构而不是什么新的或有兴趣的事物。自动地查找依赖关系可能是一种有用的方法, 这类知识可被其它模式抽取算法使用，比如可用于解释造成某种变化的原因。

⑺ 概念描述

概念描述就是对某类对象的内涵进行描述，并概括这类对象的有关特征。用户常常还需要抽象的有意义的描述。经过归纳的抽象描述能概括大量的关于类的信息。概念描述分为特征性描述和区别性描述，前者描述某类对象的共同特征，后者描述两个或更多个类对象之间的区别。

⑻ 偏差检测

偏差检测的基本方法是寻找观测结果与参照值之间有意义的差别。通过发现异常，可以引起人们对特殊情况的加倍注意。异常包括如下几种可能引起人们兴趣的模式:不满足常规类的异常例子；出现在模式边缘的特异点；与父类或兄弟类有显著不同的类；在不同时刻发生了显著变化的某个元素或集合；观察值与模型推算出的期望值之间有显著的差异的事例。偏差分析的一个重要特征就是它可以有效地过滤大量的不感兴趣的模式。

数据挖掘中常用技术

目前市面上数据挖掘应用方面有着种类繁多的商品工具和软件，大致可以归纳为下列主要类型：

⑴ 传统主观导向系统：这是针对专业领域应用的系统。如基于技术分析方法对金融市场进行分析。采用的方法从简单的走向分析直到基于高深数学基础的

分形理论和谱分析。这种技术需要有经验模型为前提。属于这类商品有美国的 Metastak、 SuperCharts、 CandlestickForecaster 和 WallStreetMoney 等。

⑵ 传统统计分析：这类技术包括相关分析、回归分析及因子分析等。一般先由用户提供假设,再由系统利用数据进行验证。缺点是需经培训后才能使用，同时在数据探索过程中，用户需要重复进行一系列操作。属于这类商品有美国的 SAS、SPSS 和 Stargraphis 等。由于近年来更先进的 DM 方法的出现和使用，这些厂商在原有系统中综合一些 DM 部件，以获得更完善的功能。以上两种技术主要基于传统的数理统计等数学的基础上，一般早已开始用于数据分析方面。

⑶ 神经元网络（NN）技术：神经元网络技术是属于软计算（Soft Computing）领域内的一种重要方法，它是多年来科研人员进行人脑神经学习机能模拟的成果，已成功地应用于各工业部门。在 DM/KDD 的应用方面，当需要复杂或不精确数据中导出概念和确定走向比较困难时，利用神经网络技术特别有效。经过训练后的 NN 可以认为具有某种专门知识的“专家”，因此可以像人一样从经验中学习。NN 有多种结构，但最常用的是多层感知机模型、反传网络、自组织映射（Self Organization Map）。它已广泛地应用于各种 DM/KDD 工具和软件中。有些是以 NN 为主导技术，例如俄罗斯的 PolyAnalyst、美国的 BrainMaker、Neurosell 和 OWL 等。NN 技术也已广泛地作为一种方法嵌入各种 DM 成套软件中。其缺点是用它来分析复杂的系统诸如金融市场，NN 就需要复杂的结构、众多的神经元以及连接数，从而使现有的事例数（不同的纪录数）无法满足训练的需要。另外由受训后的 NN 所代表的预测模型的非透明性也是其缺点。尽管如此，它还是广泛而成功地为各种金融应用分析系统所采用。神经网络的概念可参看文献[3]，贝叶斯神经网络在数据挖掘中的应用可参看文献[4]。

⑷ 决策树：在知识工程领域，决策树是一种简单的知识表示方法，它将事例逐步分类成代表不同的类别。由于分类规则是比较直观的，因而比较易于理解。这种方法一般限于分类任务。在系统中采用这种方法的有美国的 IDIS、法国的 SIPINA、英国的 Clementinc 和澳大利亚的 C5.0。有关决策树的原理和应用可参考文献[5-7]。

⑸ 基于范例的推理方法（CBR—Case based reasoning）：这种方法的思路非常简单，当预测未来情况或进行正确决策时，系统寻找与现有情况相类似的事例，并选择最佳的相同的解决方案，这种方法能用于很多问题的求解，并获得了好的结果，其缺点是系统不能生成汇总过去经验的模块或规则。采用这种方法的系统有美国的 Pattern Recognition Workbench 和法国的 KATEtools。该方法可参考有关文献[8，9]。

⑹ 进化式程序设计（Evolutionary programming）：这种方法的独特思路是系

统自动生成有关目标变量对其他多种变量依赖关系的各种假设，并形成以内部编程语言表示的程序。内部程序（假设）的产生过程是进化式的，类似于遗传算法过程。当系统找到较好地描述依赖关系的一个假设时，就对该程序进行各种不同的微小修正，生成子程序组，再在其中选择能更好地改进预测精度的子程序，如此依次进行，最后获得达到所需精度的最好程序时，由系统的专有模块将所找到的依赖关系由内部语言形式转换成易于为人们理解的明显形式，如数学公式、预测表等。由于采用通用编程语言，这种方法在原则上能保证任何一种依赖关系和算法都能用这种语言来描述。这种方法的商用产品还只见诸于俄罗斯的PolyAnalyst。据报导，它用于金融到医疗等方面的各种应用中，获得了很好的结果。

⑺ 遗传算法（Genetic Algorithms，GA）：基于进化理论，并采用遗传结合、遗传变异、以及自然选择等设计方法的优化技术。严格说来，数据分析（Data Analysis, DA）不是 GA 应用的主要领域，它是解决各种组合或优化问题的强有力的手段，但它在现代标准仪器表中也用来完成 DA 任务。这种方法的不足之处是：这种问题的生成方式使估计所得解答的统计意义的任何一种机会不再存在。另外一方面，只有专业人员才能提出染色体选择的准则和有效地进行问题的描述与生成。在系统中包含遗传算法的有美国的 GeneHunter。遗传算法的应用参考文献[7, 10]。

⑻ 非线性回归方法：这种方法的基础是在预定的函数的基础上，寻找目标度量对其它多种变量的依赖关系。该方法在金融市场或医疗诊断的应用场合，比较好地提供可信赖的结果。在俄罗斯的 PalyAnalyst 以及美国的 Neuroshell 系统中包括了这种技术。

⑼ 模糊集[11]：模糊集是表示和处理不确定性数据的重要方法。其利用隶属函数刻画不确定性，用部分代替归属的概率。不仅可以处理不完全数据、噪声或不精确数据，而且在开发数据的不确定性模型方面很有用，能提供比传统方法更灵巧、更平滑的性能。

⑽ 粗糙集：粗糙集取模糊概念之长，去隶属函数之短（需借助先验知识），成为研究模糊现象的又一有力工具。其不需要先验假设，而由集合论中的下近似和上近似来定义的。下近似中的每一个成员都是该集合的确定成员，而不是上近似中的成员肯定是该集合的成员。粗糙集的上近似是下近似和边界区的合并。边界区的成员可能是该集合的成员，但不是确定的成员。可以认为粗糙集是具有三值隶属函数的模糊集，即是、不是、也许。与模糊集一样，它是一种处理数据不确定性的数学工具，常与其它算法如规则归纳、分类和聚类方法结合起来使用，很少单独使用。有关粗糙集的原理和应用参考文献[12-14]。

⑾ 支持矢量机[15]（Support Vector Machine, SVM）：是一种基于统计学习理论的结构风险最小化的原则上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；在解决数据的分类、特征识别、图像压缩等问题方面也取得了一定进展。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。

⑿ 近邻算法[16]：将数据集合中每一个记录进行分类的方法。

⒀ 规则推导：从统计意义上对数据中的“如果-那么”规则进行寻找和推导。

⒁ 贝叶斯信念网络：其用图表示概率分布，是一种直接的、非循环的图；节点表示属性变量；边表示属性变量之间的概率依赖关系。与每个节点相关的是条件概率分布，描述该节点与它的父节点之间的关系。

⒂ 可视化：就是把数据、信息和知识转化为可视的表示形式的过程。可视化为人类和计算机这两个强大的信息处理系统提供了一个接口。使用高效的可视化界面，可以快速高效地与大量数据打交道，以发现其中隐藏的特征、关系、模式和趋势等。从而引导出新的预见和更高效的决策。

为形象直观地阐述数据挖掘的应用，图 1.5 给出了简单的两类分类问题[2]，不同的符号代表不同的类别。每一个点代表过去某一时期某银行的贷款情况，水平轴代表收入，纵轴代表全部贷款（包括抵押、买车贷款等）。分为两类：符号 x 代表那些拖欠贷款的人，符号 o 则代表讲究信誉按期付款的人。这样简单地挖掘历史数据可以帮助银行对是否贷款给某人做出抉择。下面几个图给出不同分类标准的分类情况。图 1.6 给出简单的线性分类边界，阴影部分代表没有贷款的一类。图 1.7 用了线性回归模型。图 1.8 显示简单的聚类分析将数据分为三类。数据同上图，只是数据符号变为“+”。图 1.9 用简单的收入临界值来分类。图 1.10 用非线性分类器如神经网络分类。图 1.11 用最近邻规则分类器分类。真正到实际应用中，数据的维数和点数则复杂的多，维数达百维甚至千维，点数成千上万都不止。因此就要具体问题具体分析，那种算法最适合，只有具体应用才会知道。

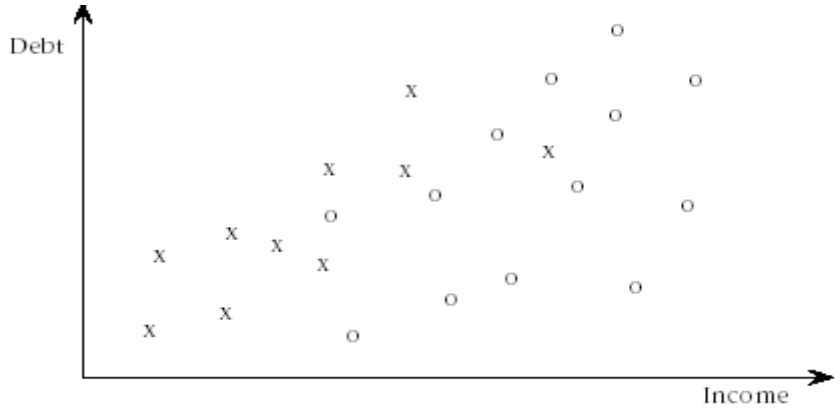


图 1.5　简单的两类分类

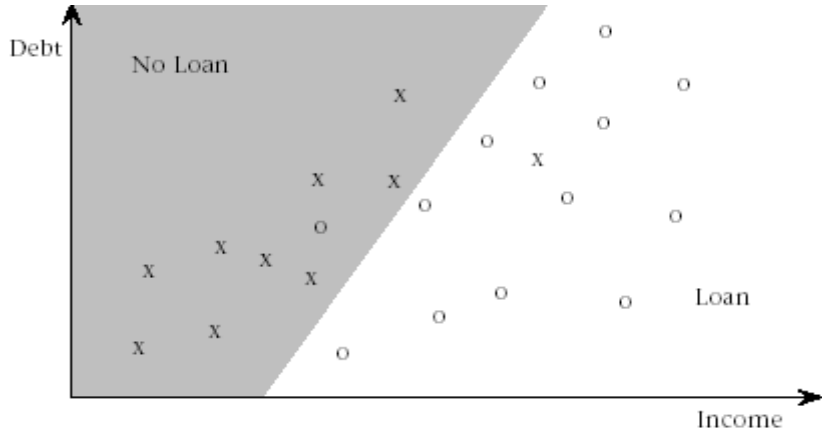


图 1.6　简单的线性分类边界，阴影部分代表没有贷款的一类

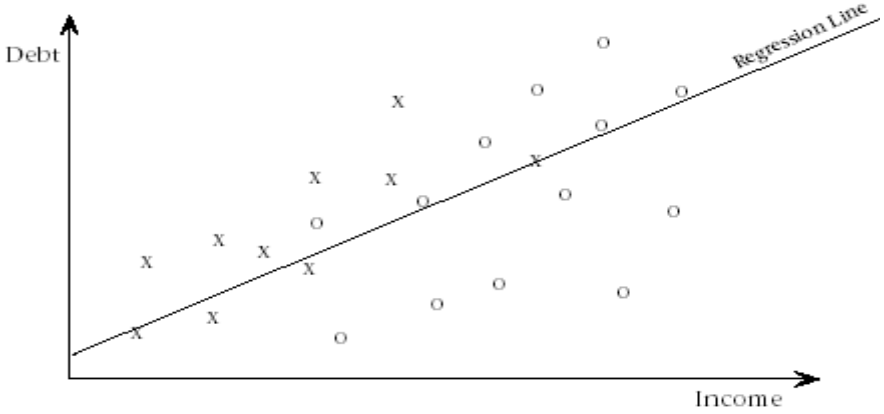


图 1.7　线性回归模型

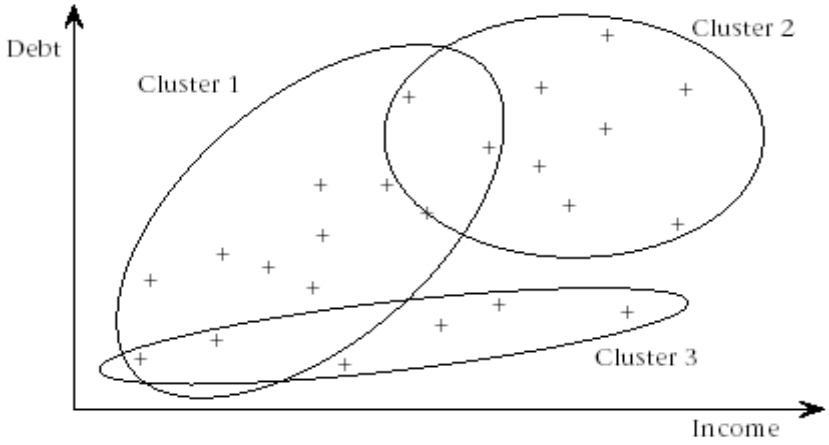


图 1.8　简单的聚类分析将数据分为三类。数据同上图，只是数据符号变为“+”

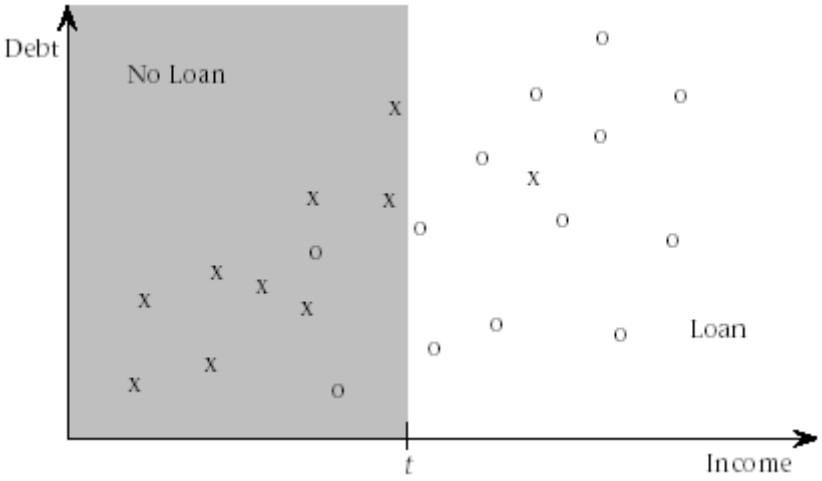


图 1.9　用简单的收入临界值来分类

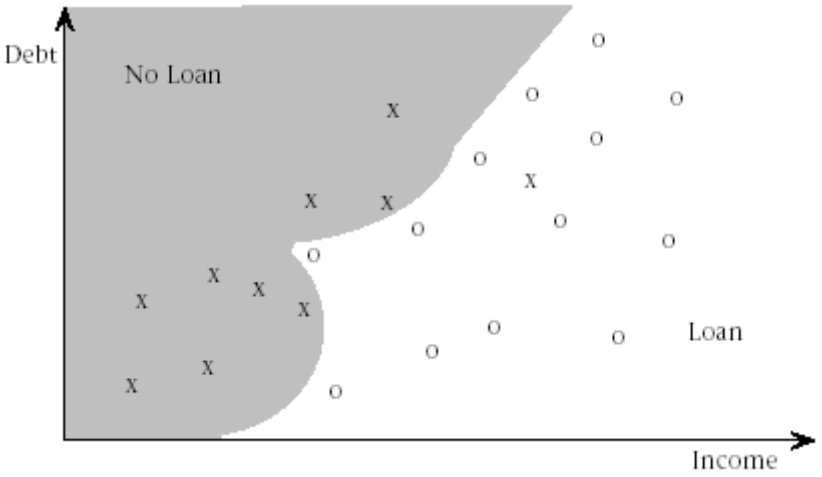


图 1.10　用非线性分类器如神经网络分类

图 1.11　用最近邻规则分类器分类

上面所列 DM 技术不可能详尽地囊括所有挖掘方法，因为多年来数理统计分析、模式识别、机器学习以及 AI 等方面的研究提供了种类繁多特点各异的方法，DM 开发人员完全可以根据不同任务加以选择使用，另外近年来在软计算（Soft Computing）和不确定信息处理（Dealing with Uncertainty of Information）方法的研究，促使 DM/KDD 技术向更深层次发展。关于数据挖掘方法的详细介绍的文章可参考文章[17-22]。

需要说明的，上面所说的 DM 中的数据是指数据库中表格形式中的记录和条目，这种数据称作结构型数据（Structured Data）。在一个企业中，还有一类像文本和网页形式的数据，称作非结构型数据(Unstructured Data)。它来自不同的信息源，如文本、图像、影视等，当然文本是最主要的一种非结构数据。1995 年分析家已预言，像文本这样非结构型数据将是在线存贮方面占支配地位的数据形式。到 1998 年初，在 Internet 上的信息网页数，已超过 5 亿。随着 Internet 的扩展和大量在线文本的出现，将标志着巨大的非结构型数据海洋中蕴藏着极其丰富的有用信息，即知识。人们从书本中获取知识的方法是阅读和理解。开发一种工具能协助用户从非结构数据中抽取关键概念以及快速而有效地检索到关心的信息，这将是一个非常引人入胜的研究领域。目前，基于图书索引检索以及超文本技术的各类搜索引擎，能协助用户寻找所需信息，但要深入发掘这类数据中的有用信息，尚需要更高层次的技术支持，如人工智能领域有关知识表示及获取的方法（如语义网络概念映射等），和自然语言理解的研究成果，可望被采用。还可能要涉及到语言学、心理学等领域。最近已出现针对文本的 DM 工具的报导，如 IBM 公司的 TexMiner、NetQuestion、WedCawler 和 megaputer 公司的 TextAnalyst 等。

采用上述技术的某些专门的分析工具已经发展了大约十年的历史，不过这些

工具所面对的数据量通常较小。而现在这些技术已经被直接集成到许多大型的工业标准的数据仓库和联机分析系统中去了。面对新经济时代，全面集成了客户、供应者以及市场信息的大型数据仓库导致公司内的信息呈爆炸性增长，企业在市场竞争中，需要及时而准确地对这些信息作复杂的分析。为了更加及时地、更加准确地做出利于企业的抉择，建立在关系数据库和联机分析技术上的数据挖掘工具为我们带来了一个新的转机。目前，数据挖掘工具正以前所未有的速度发展，并且扩大着用户群体，在未来越加激烈的市场竞争中，拥有数据挖掘技术必将比别人获得更快速的反应，赢得更多的商业机会。

数据挖掘模式的种类

(1) 分类模式　是一种分类器，能够把数据集中的数据映射到某个给定的类上，从而可以应用于数据预测。它常表现为分类树，根据数据的值从树根开始搜索，沿着数据满足的分支往上走，走到树叶就能确定类别。

(2) 回归模式　与分类模式相似，其差别在于分类模式的预测值是离散的，回归模式的预测值是连续的。

(3) 时间序列模式　根据数据随时间变化的趋势预测将来的值。其中要考虑时间的特殊性质，只有充分考虑时间因素，利用现有的数据随时间变化的一系列值，才能更好地预测将来的值。

(4) 聚类模式　把数据划分到不同的组中，组之间的差别尽可能大，组内的差别尽可能小。与分类模式不同，进行聚类前并不知道将要划分成几个组和什么样的组，也不知道根据那些数据项来定义组。

(5) 关联模式　是数据项之间的关联规则。而关联规则是描述事物之间同时出现的规律的知识模式。发现关联规则要经过以下三个步骤：①数据连接，数据准备；②给定最小支持度和最小可信度，利用数据挖掘工具提供的算法发现关联规则；③可视化显示、理解、评估关联规则。在关联规则的挖掘中要注意以下几点：

① 充分理解数据。

② 目标明确。

③ 数据准备工作要做好。能否做好数据准备又取决于前两点。数据准备将直接影响到问题的复杂度及目标的实现。

④ 选取恰当的最小支持度和最小可信度。依赖于用户对目标的估计，如果取值过小，那么会发现大量无用的规则，不但影响执行效率、浪费系统资源，而且可能把目标埋没；如果取值过大，则又有可能找不到规则，与知识失之交臂。

⑤ 很好地理解关联规则。数据挖掘工具能够发现满足条件的关联规则，但

它不能判定关联规则的实际意义。对关联规则的理解需要熟悉业务背景，丰富的业务经验对数据有足够的理解。在发现的关联规则中，可能有两个主观上认为没有多大关系的物品，它们的关联规则支持度和可信度却很高，需要根据业务知识和经验，从各个角度判断这是一个偶然现象或有其内在的合理性；反之，可能有主观上认为关系密切的物品，结果却显示它们之间相关性不强。只有很好的理解关联规则，才能去其糟粕，取其精华，充分发挥关联规则的价值。

序列模式与关联模式相似，它把数据之间的关联性与时间联系起来。为了发现序列模式，不仅需要知道事件是否发生，而且需要确定事件发生的时间。在解决实际问题时，经常要同时使用多种模式。分类模式和回归模式使用最为普遍。

数据挖掘常用的工具:

⑴ 基于神经网络的工具：由于对非线性数据的快速建模能力，神经网络很适合非线性数据和含噪声数据，所以在市场数据库的分析和建模方面应用广泛。

⑵ 基于关联规则和决策树的工具：大部分数据挖掘工具采用规则发现或决策树分类技术来发现数据模式和规则，其核心是某种归纳算法。

⑶ 基于模糊逻辑的工具：其发现方法是应用模糊逻辑进行数据查询、排序等。

⑷ 综合多种方法的工具：不少数据挖掘工具采用了多种挖掘方法，这类工具一般规模较大，适于大型数据库或者并行数据库。

数据挖掘工具的选择

在数据挖掘技术日益发展的同时，出现了许多数据挖掘工具，如何选择满足需要的数据挖掘工具已成为一个问题。具体的评价标准应从以下几方面考虑：

1. 产生的模式种类的多少

2．解决复杂问题的能力

⑴ 数据量的增大，对模式精细度、准确度要求的增高都会导致问题复杂性的增大。数据挖掘系统可以提供下列方法解决复杂问题：

⑵ 多种模式。多种模式的结合使用有助于发现有用的模式，降低问题的复杂性。例如，首先用聚类的方法把数据分组，然后再在各个组上挖掘预测性的模式，将会比单纯在整个数据集上进行操作更有效、准确度更高。

⑶ 多种算法。多种算法有很多模式，特别是与分类有关的模式，可以用不同的算法来实现，各有各的优缺点，适用于不同的需求和环境。数据挖掘系统提供多种途径产生同种模式，将更有能力解决复杂问题。

⑷ 验证方法。在评估模式时有多种可能的验证方法，比较成熟的方法像 N

层交叉验证或 Bootstrapping 等可以控制，以达到最大的准确度。

⑸ 可视化。可视化工具提供了直观、简洁的方法，方便了用户，更有助于定位重要的数据，评价模式的质量，从而减少建模的复杂性。

⑹ 数据选择和转换。数据选择和转换模式通常被大量的数据项隐藏。有些数据是冗余的，有些数据是完全无关的。而这些数据项的存在会影响到有价值的模式的发现。数据挖掘系统的一个很重要功能就是能够处理数据复杂性，提供挖掘工具，选择正确的数据项和转换数据值。

⑺ 扩展性。为了更有效的提高处理大量数据的效率，数据挖掘系统的扩展性十分重要。要了解数据挖掘系统能否充分利用硬件资源？是否支持并行性能？支持那种并行计算机？当处理器的数量增加时，计算规模是否相应增长？是否支持数据并行存储？为单处理器的计算机编写的数据挖掘算法不会在并行计算机上自动以更快的速度运行。为充分发挥并行计算的优点，需要编写支持并行计算的算法。

⑻ 易操作性。操作性能的好坏是一个至关重要的因素。图形界面友好的工具可以方便用户，引导用户执行任务，为用户节省时间。提供嵌入技术的工具更是它的可取之处，通过嵌入到应用程序中，缩短了开发时间。既可以将模式运用到已存在或新增加的数据上，也可以把模式导出到程序或数据库中。例如：有的工具有图形化界面，引导用户半自动化地执行任务，有的使用脚本语言。有些工具还提供数据挖掘的 API，可以嵌入到像 C、VisualBasic、PowerBuilder 这样的编程语言中。有的允许通过使用 C 这样的程序语言或 SQL 中的规则集，把模式导出到程序或数据库中。

⑼ 数据存取能力。好的数据挖掘工具可以使用 SQL 语句直接从数据库管理系统 DBMS 中读取数据。这样可以简化数据准备工作，并且可以充分利用数据库的优点。没有一种工具可以支持大量的 DBMS，但可以通过通用的接口连接大多数流行的 DBMS。Microsoft 的 ODBC 就是一个这样的接口。

⑽ 与其他产品的接口。传统的查询工具、可视化工具可以帮助用户理解数据和结果。这些工具可以是传统的查询工具、可视化工具、OLAP 工具。数据挖掘工具能否提供与这些工具集成的简易途径是衡量数据挖掘工具好坏的标准。

通过对数据挖掘种类的分析，给出了数据挖掘工具的选择标准。因为数据挖掘工具需要考虑的因素很多，很难按照原则给工具排一个优劣次序。最重要的还是用户的需要，根据特定的需求加以选择，文中考虑的因素仅为充分利用数据挖掘工具提供参考。数据挖掘工具可以给很多产业带来收益。国外的许多行业如通信、信用卡公司、银行和股票交易所、保险公司、广告公司、商店等已经大量利用数据挖掘工具来协助其业务活动，国内在这方面的应用还处于起步阶段，对数

据挖掘技术和工具的研究人员以及开发商来说，我国是一个有巨大潜力的市场。

数据挖掘的范围

追根溯源，“数据挖掘”这个名字来源于它有点类似于在山脉中挖掘有价值的矿藏。在商业应用里，它就表现为在大型数据库里面搜索有价值的商业信息。这两种过程都需要对巨量的材料进行详细地过滤，并且需要智能且精确地定位潜在价值的所在。对于给定了大小的数据库，数据挖掘技术可以用如下的超能力产生巨大的商业机会：

⑴ 自动趋势预测。数据挖掘能自动地在大型数据库里寻找潜在的预测信息，对将来的趋势和行为进行预测，从而很好地支持人们的决策。传统上需要很多专家来进行分析的问题，现在可以快速而直接地从数据中找到答案。一个典型的利用数据挖掘进行预测的例子就是目标营销。数据挖掘工具可以根据过去邮件推销中的大量数据找出其中最有可能对将来的邮件推销做出反应的客户。有些数据挖掘工具还能够解决一些很消耗人工时间的传统问题，因为它们能够快速地浏览整个数据库，找出一些专家们不易察觉的极有用的信息。

⑵ 探测以前未发现的模式。数据挖掘工具扫描整个数据库并辨认出那些隐藏着的模式，比如通过分析零售数据来辨别出表面上看起来没联系的产品，实际上有很多情况下是一起被售出的情况。

⑶ 数据挖掘技术可以让现有的软件和硬件更加自动化，并且可以在升级的或者新开发的平台上执行。当数据挖掘工具运行于高性能的并行处理系统上的时候，它能在数分钟内分析一个超大型的数据库。这种更快的处理速度意味着用户有更多的机会来分析数据，让分析的结果更加准确可靠，并且易于理解。

⑷ 此外，数据库可以由此拓展深度和广度。深度上，允许有更多的列存在。以往，在进行较复杂的数据分析时，专家们限于时间因素，不得不对参加运算的变量的数量加以限制，但是那些被丢弃而没有参加运算的变量有可能包含着另一些不为人知的有用信息。现在，高性能的数据挖掘工具让用户对数据库能进行通盘的深度编历，并且任何可能参选的变量都被考虑进去，再不需要选择变量的子集来进行运算了。广度上，允许有更多的行存在。更大的样本让产生错误和变化的概率降低，这样用户就能更加精确地推导出一些虽小但颇为重要的结论。

数据挖掘的体系结构

现有很多数据挖掘工具是独立于数据仓库以外的，它们需要独立地输入输出数据，以及进行相对独立的数据分析。为了最大限度地发挥数据挖掘工具的潜力，它们必须象很多商业分析软件一样，紧密地和数据仓库集成起来。这样，在人们

对参数和分析深度进行变化的时候，高集成度就能大大地简化数据挖掘过程。

集成后的数据挖掘体系有自己的特点。应用数据挖掘技术，较为理想的起点就是从一个数据仓库开始，这个数据仓库里面应保存着所有客户的合同信息，并且还应有相应的市场竞争对手的相关数据。这样的数据库可以是各种市场上的数据库如 Sybase、Oracle、Redbrick 等等，并且可以针对其中的数据进行速度上和灵活性上的优化。

联机分析系统(OLAP)服务器可以使一个十分复杂的最终用户商业模型应用于数据仓库中。数据库的多维结构可以让用户从不同角度，--比如产品分类、地域分类、或者其他关键角度--来分析和观察他们的生意运营状况。数据挖掘服务器在这种情况下必须与联机分析服务器，以及数据仓库紧密地集成起来，这样就可以直接跟踪数据和并辅助用户快速作出商业决策，并且用户还可以在更新数据的时候不断发现更好的行为模式，并将其运用于未来的决策当中。

数据挖掘系统的出现代表着常规决策支持系统的基础结构的转变。不象查询和报表语言仅仅是将数据查询结果反馈给终端用户那样，数据挖掘高级分析服务器把用户的商业模型直接应用于其数据仓库之上，并且反馈给用户一个相关信息的分析结果。这个结果是一个经过分析和抽象的动态视图层，通常会根据用户的不同需求而变化。基于这个视图，各种报表工具和可视化工具就可以将分析结果展现在用户面前，以帮助用户计划将采取怎样的行动。

DM/KDD 面临的问题

DM/KDD 技术发展较快，理论也在不断成熟，但随着研究日益深入，发展速度也越来越慢，主要是 DM/KDD 所使用的各项技术理论发展有波折。神经网络理论几十年来突破不大，关于分类、决策树等的新进展不多，这些都是 DM/KDD 到目前尚不能令人满意的原因。况且 DM/KDD 的许多技术源于机器学习方法，但由于现实世界数据库存在一些固有的特点，因此给 DM/KDD 带来一些难点。正是这些关键之处，才形成了 DM/KDD 领域自己独特的研究方向。有关 DM/KDD 的研究和应用还面临着一部分比较突出的问题[2]。

⑴ 超大数据量和数据库问题　数据库中数据的迅速增长是 DM/KDD 得以发展的原因之一，这也正是对 DM/KDD 研究的挑战。目前，含有几百个域和表，具有上千万条记录和数据规模的数据库已经十分常见，连万亿（10^{12}）级字节的数据库也出现了。穷举法、经验分析方法对数兆字节、数千兆字节甚至数太拉字节的数据显得无能为力。此时 DM/KDD 系统必须采用一定的数据汇集方法，根据用户定义的发现任务，选择有关的域空间，采取随机抽样的方法，对样本进行分析。处理大规模数据的方法还需要研究和开发，这些方法包括有效算法、近似方

法及大规模并行处理算法等。

⑵ 数据的高维问题　现在的数据库不仅有大量的记录，通常还有很多字段(如各种特性)，因此导致数据的维数过高。高维数据增加了模型搜索空间的规模，同时也增加了算法找到无用模式的可能性。解决此类问题的方法是减少问题的有效维数，使用一些方法（如主分量分析方法）识别不相关的变量。

⑶ 过分拟合（overfitting）问题　当算法在有限数据集中搜索某一模型的最优参数时，建立了一般模式，但其中包括了该数据集的噪声影响，这就使得模型应用于其他试验数据时会失效。这类问题可以通过交叉确认、重整化或其他高级统计方法尝试解决。

⑷ 数据的变化、稀疏、不完整、含有噪声、存在冗余问题　这个问题在商业数据库中尤为尖锐，从数据库中提取规则、发现知识时，往往会发现数据变化、或稀疏、或缺少有关数据、亦或存在许多无用的数据。如果数据库预先没有设计好，就可能遗失重要的特征。解决方法一般是使用专门技术及高级统计方法来识别隐含的变量间的相关性。

① 动态变化的数据

数据的动态变化是大多数数据库的一个主要特点。一个联机系统应能够保证数据的变化不会导致错误的发现。

② 噪声

由于人为因素的影响，如数据的手工录入以及主观选取数据等，从而使得数据具有噪声。带噪声的数据会影响抽取的模式的准确性。

③ 数据不完整

数据库中某些个别的记录，其属性域可能存在空值现象，另外对某一发现来说还可能完全不存在其所必需的记录域。这种数据的不完整性将给发现、评估和解释一些重要的模式带来困难。

④ 冗余信息

数据库中同一信息有时存储在多个地方。函数依赖就是一个通常的冗余形式。冗余信息可能造成错误的知识发现,至少有些发现是用户完全不感兴趣的。为避免这种情况发生,系统需要知道数据库中有哪些固有的依赖关系。

⑤ 数据稀疏

相应于可能的巨大的发现空间,数据库中所记录的实际数据的密度是非常稀疏的。这对传统的经验定律发现方法是个挑战。

⑸ 域之间的复杂联系问题　一个数据库的分层建构的属性或变量、属性之间的关系，以及表达知识的更高级方法，需要可以有效地统一这些信息的算法。一般 DM/KDD 算法针对的是简单属性值记录，针对域之间复杂的联系，目前开发

的一些方法都还不十分成熟。

⑹ 用户交互和先验知识问题 很多现有的 DM/KDD 方法和工具并不是真正交互的（除了一些简单方法），也不可能纳入已有知识。在 DM/KDD 过程的每一步骤中，了解区域的情况是重要的。如贝叶斯方法把数据和分布的先验概率作为嵌入的先验知识。有些方法用数据库的演绎能力发现知识，然后用这些知识指导数据挖掘。

⑺ 模式泛滥问题 许多 DM/KDD 模型很容易找到模式，但是明显冗余，如“人都有两只手”。减少这种明显的“发现”的一般方法是将焦点集中于变化上，因为这种模式不会随着提取规则的不同而变化。冗余的发现还可以通过规则求精方法消除。而更困难的任务是将重要的模式从无用的专业知识中分离出来。

⑻ 其他问题 与现成系统的合成问题：一个孤立的发现系统用途不大，即使建立得很完善，仍需要与其他现成的硬件/软件系统合成，如 DM/KDD 通过询问界面与 DBMS 的结合等。

以上是现实世界数据库中存在的一些不利因素。在 DM/KDD 发展的道路上，还有许多困难要加以克服，有许多问题有待研究，如不适当的统计知识、过多的冗余模式、现有系统的集成、多策略系统等等。

数据挖掘实际应用

DM/KDD 工具和软件已在各个部门得到很好的应用，并收到明显的效益。

⑴ 金融方面：银行信用卡和保险行业，预测存/贷款趋势，优化存/贷款策略，用 DM 将市场分成有意义的群组和部门，从而协助市场经理和业务执行人员更好地集中于有促进作用的活动和设计新的市场运动。

⑵ 在客户关系管理方面：DM 能找出产品使用模式或协助了解客户行为，从而可以改进通道管理（如银行分支和 ATM 等）。又如正确时间销售（Right Time Marketing）就是基于顾客生活周期模型来实施的。

⑶ 在零售业/市场营销方面：是数据挖掘技术应用最早也是最重要的领域，DM 用于顾客购货篮的分析可以协助货架布置，促销活动时间，促销商品组合以及了解滞销和畅销商品状况等商业活动。通过对一种厂家商品在各连锁店的市场共享分析，客户统计以及历史状况的分析，可以确定销售和广告业务的有效性。

⑷ 在过程控制/质量监督保证方面：DM 协助管理大量变量之间的相互作用，DM 能自动发现某些不正常的数据分布，暴露制造和装配操作过程中变化情况和各种因素，从而协助质量工程师很快地注意到问题发生的范围并采取相应的改正措施。

⑸ 在远程通讯部门：基于 DM 的分析协助组织策略变更以适应外部世界的

变化，确定市场变化模式以指导销售计划。在网络容量利用方面，DM 能提供对客户组类服务使用的结构和模式的了解，从而指导容量计划人员对网络设施做出最佳投资决策。

⑹ 化学/制药行业：从各种文献资料中自动抽取有关化学反应的信息，发现新的有用化学成分。

⑺ 遥感领域：针对每天从卫星上及其它方面来的海量数据，对气象预报、臭氧层监测等能起很大作用。

⑻ 军事方面：使用 DM 进行军事信息系统中的目标特征提取、态势关联规则挖掘等。

⑼ 天文方面：使用 DM 对天体分类、挖掘特殊的、稀有的或新的天体或天文现象等。

总而言之，数据挖掘和知识发现（DM/KDD）技术被广泛应用于各行各业：股市预测、银行金融、零售与批发、制造、保险、公共设施、政府、教育、远程通讯、软件开发、运输、临床数据分析、医学诊断、信息查找、控制算法采集和过程控制、复杂化学化合物分析、结构工程、市场分析、经济投资决策等各个企事业单位及国防科研上。而较具代表性的科研发现的例子是：天文学方面，对天体的分类、金星上火山的发现等；生命科学方面，由 DNA 库确定基因及遗传信息；地球科学方面，对地壳和板块的监测从而预报地震等；有名的 BACON 系统（用于由数据库数据产生反映数据规律的代数方程的系统）就十分成功地对开普勒定律进行了重新发现。据报导，DM 的投资回报率有达 400%甚至 10 倍的事例。数据挖掘的典型应用是在商业领域，但其方法和技术能否应用于其它领域，现在似乎已有突破，如将其应用于医疗、天文学、地学、生物学等领域。在这里提出了问题，但没有论及在不同领域中的具体应用，原因是知识的局限，希望能将数据仓库、数据挖掘等技术应用于科学数据库，从而丰富科学数据库的内容，并将科学数据库的应用推向新的深度。

动态与展望

以上从数据挖掘和知识发现的概貌、背景、兴起的原因、研究内容、研究方法和关键问题及典型系统等方方面面，对其作了一个详细的综述。由于篇幅所限，还有许多问题未涉及，如可听技术的应用。数据挖掘和知识发现的研究正方兴未艾具有非常广阔的前景。如利用粗糙集(rough set)作为 DM/KDD 的工具，面向多数据库的 DM/KDD，文本数据库中的知识发现，贝叶斯网络模型的使用，面向多策略和合作的发现系统，面向对象的 DM/KDD，结合多媒体技术的应用等等都是新的研究方向。目前，DM/KDD 研究的重点，正从理论转向应用，可说凡

是用到数据库的地方，就有 DM/KDD 的课题等待人们去探讨。我们准备在天文的 DM/KDD 应用方面做一些探索，为促进 DM/KDD 在天文方面的应用做一点尝试。DM/KDD 与科学数据库的结合对科技的发展必定会起到很大的促进作用，具有广泛的应用前景。

参 考 文 献

[1] Fayyad U, Piatetsky-Shapiro G., Smyth P, 1996, From Data Mining to Knowledge Discovery in Databases: An Overview. In Advances in Knowledge Discovery and Data Mining, eds. U Fayyad, G Piatetsky-Shapiro, P Smyth, and R Uthurusamy, 1-30. Menlo Park, Calif.: AAAI Press

[2] Fayyad U, Piatetsky-Shapiro G, Smyth P, "From Data Mining to Knowledge Discovery in Databases", http://www.kdnuggets.com/publications/surveys.html

[3] Haykin S, 1994, Neural Networks: A Comprehensive Foundation, Macmillan/IEEE Press

[4] Heckerman D, 1997, Bayesian networks for data mining, Data Mining and Knowledge Discovery, 1: 79-119

[5] Quinlan J R, 1987, Simplifying Decision Trees, Internet. Journal of Man-Machine Studies, 27: 21-234

[6] Quinlan J R, 1987, Generating production rules from decision trees, Proceedings of IJCAI-87, Milan, Italy

[7] Quinlan J R, 1988, An empirical comparison of genetic and decision-tree classifiers, Proceedings of ICML-88, San Mateo, CA

[8] Kolodner J L, 1993, Cased-based Reasoning, Morgan Kaufmann

[9] Aamodt A, & Plaza E, 1994, Cased-based reasoning: foundational issues, methodological variations, and system approaches, AI Communications, 7(1): 39-59

[10] Goldberg D E, 1989, Genetic Algorithms in Search, Optimization, and Machine Learning, Addison-Wesley

[11] Zadeh L A, 1965, Fuzzy sets. Information and Control, 8: 338-353

[12] Grzymala-Busse J W, Ziarko W, 2000, Data Mining and Rough Set Theory, CACM, 43: 108-109

[13] Pawlak Z, 1982, Rough Sets, International J of Computer and Information

Sciences, (11): 341-356

[14] Lin Tsau Young, 1997, Rough Sets and Data Mining: Analysis of Imprecise Data, Kluwer Academic Publishers

[15] Burges J C J, 1998, A tutorial on support vector machines for pattern recognition, Data Mining and Knowledge Discovery, 2(2): 121-167

[16] Dasarathy B V, 1991, Nearest Neighbor (NN) Norms: NN Pattern Classification Techniques, Los Alamito, CA: IEEE Computer Society Press

[17] Goebel M, & Gruenwald L, A Survey of Knowledge Discovery and Data Mining Tools. Technical Report, University of Oklahoma, School of Computer Science, Norman, OK, February 1998.

[18] 史忠植，“知识发现”，清华大学出版社，2002

[19] 边肇祺，张学工等，“模式识别”，清华大学出版社，1999

[20] Berry M, & Linoff G, 1997, Data Mining Techniques, John Wiley

[21] Michalski R S, Kaufman K A, 1997, Data Mining and Knowledge Discovery: A Review of Issues and A Multistrategy Approach, Machine Learning and Data Mining: Methods and Applications, John & Sons Ltd, 92-107

[22] Han Jiawei, Micheline Kamber, 2000, Data Mining: Concepts and Techniques, Morgan Kaufmann Publishers

§1.3.2 天文中的数据挖掘和知识发现

本节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、过程和任务。分析了当前天文数据的复杂性特征，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在天文数据挖掘和知识发现领域的研究进展及其热点，并阐述了其所面临的挑战。天文中的数据挖掘和知识发现的出现和发展，预示其在二十一世纪将具有广泛的应用前景，将对天文学的发展起到巨大的推动作用，并在知识和技术等方面对天文学家提出了新的挑战。而且，数据挖掘技术在虚拟天文台中的成功应用，是虚拟天文台充分发挥作用的关键所在。

天文中的数据挖掘和知识发现的兴起

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，各个领域正面临着一场“数据爆炸”，即数据量呈指数增长。在未来十年里，将产生比过去所有数据总和还要多的数据。尽管目前分析和处理数据的方法和技术远远滞后于数据的增长，人们已逐步意识到这些数据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、探测器效率呈摩尔规律增长、巡天技术的发展，都给天文学带来革命性的变化：数据通常以 TB，甚至 PB 计量。更多的地面和空间天文设备、以及更大口径和更精密仪器的投入使用，将带来天文数据的进一步飞速增长，例如哈勃空间望远镜每天大约产生 5GB 的数据，筹建中的大口径综合巡天望远镜(Large-Aperture Synoptic Survey)日产数据将高达 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是虚拟天文台成功的重要因素[2-3]。

随着计算机技术、数据库技术、统计学、数学、机器学习等方面在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。知识发现就是对数据抽取和精化而取得新知识的模式，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，关系型数据库应用广泛，并且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。因此，数据库知识发现(Knowledge Discovery in Database， KDD)的研究非常活跃。该术语最早由 Fayyad 于 1989 年提出，并定义 KDD 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 等人的文章[4]。具体到天文学中，数据挖掘和知识发现（Astronomical Data Mining 和 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或现象，或者发现以前未知种类的天体或新天文现象。近年来，这方面的研究已成为天文数据研究领域的热点。

数据挖掘和知识发现过程

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

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，它是根据用户的需要从原始数据库抽取一组数据。数据预处理一般包括消除噪声、推导计算缺值数据、消除重复记录、完成数据类型转换等。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一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。任务确定后，要选择挖掘算法。同样的任务可用不同的算法实现，选择算法时要考虑两个因素：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要看用户或实际运行系统的要求，有的希望获得描述性的、容易理解的知识，而有的则是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，但要获得好的挖掘效果，必须对各种挖掘算法的要求或前提假设有充分理解。

当挖掘结束时，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。为了使人们更好地理解结果，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：

⑴ 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，则挖掘的结果就不会好。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。

⑵ 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结果，这就需要重复前面的过

程，甚至从头重新开始。

⑶ 可视化技术在数据挖掘的各个阶段扮演着重要的角色。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，从而可以更好地选择数据，如去掉一些极大和极小的离群数据；在数据挖掘阶段，通过可视化技术可对挖掘过程有一个直观的了解，从而可以控制挖掘过程；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结果，对结果做出合理的理论解释。天文中的可视化工具有 XGobi、ExplorN、ViSta、CViz、IVEE 和其它的一些软件包，可以提供 2 维或多维数据浏览、平行坐标图等功能[5]。

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据着重要地位。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）及 Kohonen 学习矢量量化方法(Learning Vector Quantization, LVQ)；另外，最近又兴起了一种新的方法—粗糙集(Rough Set)。

聚类是根据数据的不同特征，将其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有兴趣的事物，这些知识可被其它模式抽取算法使用。经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的或奇异的天体，如褐矮星和高红移类星体的发现。

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所决定的，并成为天文数据挖掘研究首要解决的问题。

⑴ 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。数据的形态有数字、符号、图形、图像等；数据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，天体才具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由此更为复杂化，从而使天文数据与其它类型数据的挖掘方法存在着差异。

⑵ 天文数据的复杂性

近年来随着天文观测技术的飞速发展，天文数据具备以下几个方面的复杂性：

① 海量数据

天文数据将以 TB 甚至 PB 计量，如此大的数据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，并发展新的高效算法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天文系统复杂性的重要标志，其中蕴含着系统内部作用的复杂机制，因而被作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，例如由于空间天文学的飞速发展，覆盖的波段的数目也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。

⑶ 天文学中的数据挖掘的科学要求

数据挖掘利用复杂的技术建立模型，从数据中发现模型和相关性。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。事件性数据挖掘进一步分为四类：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着人们感兴趣的已知现象，无论空间上或时间上；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着预期的相关性来预测数据中存在着以前未知的事件；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：在天空中的同一位置证认天体；②时间相关：证认发生在相同时间或相关时间的事件或现象；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，天文中的数据挖掘的科学要求有如下几种[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，用来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，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(the Palomar Digital Sky Survey)和SDSS(the Sloan Digital Sky Survey)巡天中，通过在双色图中远离正常恒星区的特性发现高红移类星体。

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）数据中发现的“bumpers”。

天文中的数据挖掘技术

⑴ 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、强大的多处理计算机、数据挖掘算法。因此要想提高算法效率，须从这三个基础做起：

① 正在建设中的虚拟天文台将把空间与地面观测设备得到的多波段巡天的海量数据有机地联合起来，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，分析探索这些数据将得到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就是采用

了并行策略，由此大幅度提高了计算效率。提高数据库查询语言的效率，大型分布式的数据库不仅要实现数据的一体化存储，解决数据的索引、组织和分布管理的问题，还需要有一体化的查询语言作为操纵的接口，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，从而减小运算的复杂度。Auton 实验室小组提出了一个方法 CSS(Cached Sufficient Statistics)能自动地从大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发展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)引进天文学。关于错误发现率在天文中的应用可参考 Miller 等人的文章[12-13]。Komarek 和 Moore 将静态的 AD 树从结构上调整为动态的 AD 树，克服了静态的三个弱点[14]。其中 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 扩展 K 均值为 X 均值，可以有效地估计数据中的类别数[16]。Moore 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 算法为基础的聚类算法的复杂度[17]。Maino 等人在独立分量分析(Independent Component Analysis, ICA)方法基础上发展了一个新而快速的算法 FastICA，用以分离天体物理参量[18]。

⑵ 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]。其特点是利用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用于处理天文数据的非线性复杂关系，并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，因而被广泛地应用于天文数据挖掘和知识发现的研究中，并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，自动地提取特征、主分量分析、聚类、编码、特征映射，有助于可视化，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]。神经网络在天文中有广泛的应用，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也可以用这种方法（如寻找高红移类星体）。NExt(Neural Extractor)是建立在神经网络基础上的软件包，可以自动地从天文图像中提取星表、寻找特殊天体、恒星/星系分类[32]。

⑶ 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可能的规律性。常用的统计方法有回归分析（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 Babu 和 Feigelson 的文章[33-34]，大多数多变量方法的 Fortran 程序可参考 Murtagh 和 Heck 的程序[35]。

EM 算法(Expectation Maximization，EM)是一种聚类算法[36-37]，在天文中常用于两种情形密度估计：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的大小范围特征，并且提供了一种证认多维色空间中远离正常恒星的天体的方法，例如高红移类星体的证认。

主分量分析方法（Principle Component Analysis, PCA）[38-40] 具有非监督性，是线性分析，具有降维去噪的功能，因而常用于对数据进行预处理，去掉一些无关或不重要的参量。在天文中有重要的应用，主要用于恒星、星系和类星体的光谱分类；星系的形态分类；自动的红移确定；通过将发射线分解为几个独立量来研究发射区的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取得了一定进展。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，例如可以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 的文章[42]。

⑷ 模糊集

对于天文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。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。如 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分类。

⑸ 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，然后再实施其他挖掘算法；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)，Roweis 和 Saul[46]提出了 LLE(locally linear embedding)都用于处理非线性的高维数据。除此之外，降维还可通过使用粗糙集理论精简维数而得以实现；而支持向量机则能够应付处理数据因维数过高而产生的复杂性。

天文中常见问题及其处理

在天文中会遇到各种各样的问题，面对这些问题如何处理，这是摆在天文学家面前的重要课题。下面的表 1.2 中列出了一些天文中经常遇到的问题及其处理方法：

表 1.2　天文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	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[43，47] 神经网络(NN)[23-31，50] 小波变换[48] 决策树[24]
图像分类	数字底片巡天中的恒星/星系区别	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[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] 傅立叶拟合[62] 变像素线性重建[63] 最大熵方法（MEM）[64] Massive Inference[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[18]

上面仅对天文中的部分问题及处理方法作了一下总结，但天文中远远不止这

些问题，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。具体问题需要具体分析，任何算法都有其优劣，对这个问题适用而对另一个问题则可能无效，或者几种算法处理问题的效果相当，而且有时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]对比了自组织映射、学习矢量量化及利用模糊分类器和 BP 神经网络的混合算法在大规模巡天中恒星与星系的自动分类中的作用，得出结论：好的分类结果依赖于算法的复杂性和可获得的训练样本。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 是模型独立的，但 IB 监督的波长群在概念上接近 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、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，并且以在天文和非天文中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 等人的文章[69]。Kim 等人[70]对比了三种聚类算法(the matched filter, MF；the adaptive matched filter, AMF；a 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。各有优点，将二者混合起来的模型称 HMF，此时 HMF 在背景均匀时优于 VTT，而当背景非均匀时前者较后者敏感，而且他们发现当对 SDSS 巡天选择合适的探测阈值时，两种算法效果相当。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 是一个很好的方法。

天文学中数据挖掘技术所面临的挑战

⑴ 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。

⑵ 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。

⑶ 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。

⑷ 提高数据挖掘方法的容易度：这包括数据挖掘自动化程度的提高；提高用户界面便以支持随机用户的浏览；提高大型分布数据的可视化程度；发展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，提供对大型的、分布式的数据挖掘的恰当的实验场所。数据挖掘技术的发展仅靠天文学家是远远不够的，这需要来自计算机界、统计学界、数学界的科学家的精诚合作；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。

结语与展望

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和质量与算法的优劣。天文数据本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。对于其他领域（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充分提高。未来的虚拟天文台将分布的海量数据有机地结合起来，同时需要发展与之匹配的强而有效的数据挖掘工具，用以处理这些高度精确的海量数据，有效地处理天文学中的“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，作为天文学家即要懂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，以适应时代发展的需要。

参 考 文 献

[1] Szalay A S, Brunner R J. In: Brian J McLean, Daniel A Golombek, Jeffrey J E Hayes, and Harry E Payne, eds. New Horizons from Multi-Wavelength Sky Surveys, Proc. Of IAU colloq. No. 179, Baltimore, 1996, s.l. 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, In: Brunner R J, Djorgovski S G, & Szalay A, eds. Virtual Observatories of the Future, ASP Conference Series, No. 225, California, Pasadena, 2000, California, San Francisco: ASP, 2001: 272

[6] Borne K D, In: Banday A J, Zaroubi S, &Bartelmann M eds. Mining the Sky, Proc. of the MPA/ESO/MPE Workshop, Germany, Garching, 2000, Heidelberg: Springer-Verlag, 2001: 671

[7] Thakar A R, Kunszt P Z, Szalay A S, In: Brunner R J, Djorgovski S G, & Szalay A, eds. Virtual Observatories of the Future, ASP Conference Series, No. 225, California, Pasadena, 2000, California, San Francisco: ASP, 2001: 230

[8] Moore R W, In: Brunner R J, Djorgovski S G, & Szalay A, eds. Virtual Observatories of the Future, ASP Conference Series, No. 225, California, Pasadena, 2000, California, San Francisco: ASP, 2001: 257

[9] Williams R, In: Brunner R J, Djorgovski S G, & Szalay A, eds. Virtual Observatories of the Future, ASP Conference Series, No. 225, California, Pasadena, 2000, California, San Francisco: ASP, 2001: 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. Mining the Sky, Proc. of the MPA/ESO/MPE Workshop, Germany, Garching, 2000, 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, astro-ph/0108362

[19] Bishop C M, Neural Networks for Pattern Recognition, Oxford UK: Oxford University Press, 1995: 1

[20] Kohonen T, The self-organizing maps, Proceedings of the IEEE, 1990, 78(9): 1464

[21] Bazell D, & Peng Y, Ap. JS, 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*. The SKICAT System for Processing and Analysing Digital Imaging Sky Surveys, Publ. Astron. Soc. Pacific, 1995, 107: 1243

[24] Weir N, Fayyad U, Djorgovski S G, Automated Star/Galaxy Classification for Digitized POSS-II, Astron. J. 1995, 109: 2401

[24] Fayyad U, Smyth P, Weir N, *et al*. Automated Analysis and Exploration of Image Databases: Results, Progress, and Challenges, J. Intel. Inf. Sys. 1995, 4: 7

[25] Bertin E, Arnout S, Astron. Astrophys. Suppl., 1996, 117: 393

[26] Storrie-Lombardi M C, Lahav O, Sodre L jr, *et al*. M.N.R.A.S., 1992, 259: 8

[27] Lahav O, Naim A, Sodre L jr, Storrie-Lombardi M C, M.N.R.A.S., 1996, 283: 207

[28] Bailer-Jones C A L, Irwin M, von Hippel T, M.N.R.A.S., 1998, 298: 361

[29] Allende Prieto C, Rebolo R, Lopez R J G, *et al*. Astron. J., 2000, 120: 1516

[30] Weaver W B, Ap. J., 2000, 541: 298

[31] Longo G, Tagliaferri R, Andreon S, In: Banday A J, Zaroubi S, &Bartelmann

M eds. Mining the Sky, Proc. of the MPA/ESO/MPE Workshop, Germany, Garching, 2000, Heidelberg: Springer-Verlag, 2001: 379

[32] Babu G J, & Feigelson E D, Astrostatistics (London: Chapman & Hall), 1996: 11

[33] Babu G J, & Feigelson E D, Statistical Challenges in Modern Astronomy II (New York: Springer-Verlag), 1997: 3

[34] Murtagh F, & Heck A, Multivariate Data Analysis, Kluwer, Dordrecht, 1987 (ISBN 90 277 2425 3, ISBN 90 277 2426 1)

[35] Nichol R C, Connolly A J, Moore A W, *et al*. In: Brunner R J, Djorgovski S G, & Szalay A eds. Virtual Observatories of the Future, ASP Conference Series, No. 225, California, Pasadena, 2000, California, San Francisco: ASP, 2001: 265

[36] Connolly A J, Genovese C, Moore A W, *et al*. 2000, astro-ph/0008187

[37] Adanti S, Battinelli P, Capuzzo-Dolcetta R, *et al*. Astron. Astrophys. Suppl. Ser. 1994, 108: 395

[38] Connolly A J, Szalay A S, Bershady M A, *et al*. Astron. J. 1995, 110(3): 1071

[39] Connolly A J, Szalay A S, Astron. J. 1999, 117: 2052

[40] Zhang Yanxia, Zhao Yongheng, PASP, 2003, 115: 1006

[41] Burges C J C, Data Mining and Knowledge Discovery, 1998, 2: 121

[42] Spiekermann G, Ap.J., 1992, 103: 2102

[43] Mahonen P, Frantti T, Ap. J., 2000, 541: 261

[44] Tenenbaum J B, de Silva V, Langford J C, Science, 2000, 290: 2319

[45] Roweis S, Saul L K, Science, 2000, 290: 2323

[46] Cortiglioni F, Mahonen P, Hakala P, *et al*. Ap. J., 2001, 556: 937

[47] Connolly A J, Castander F, Genovese C, Hilton E, et al. In: Banday A J, Zaroubi S, &Bartelmann M eds. Mining the Sky, Proc. of the MPA/ESO/MPE Workshop, Germany, Garching, 2000, Heidelberg: Springer-Verlag, 2001: 323

[48] Naim A, Ratnatunga K U, Griffiths E, Ap. JS., 1997, 111: 357

[49] Naim A, Lahav O, Sodre L Jr, and Storrie-Lombardi M C, M.N.R.A.S, 1995, 275: 567

[50] Mahonen P, Hakala P J, Ap. J., 1995, 452: 77

[51] Mahonen P, & Frantti T, Ap. J., 2000, 541: 261

[52] Murtagh F D, in “Astronomical Data Analysis Software and Systems I”, ASP Conf. Series, 1992, 25: 265

[53] Jarvis J F, Tyson J A, in “Instrumentation in Astronomy III”, SPIE, 1979, 172: 422

[54] Jarrett T H, Chester T, Cutri R, Schneider S, Skrutskie M, Huchra J P, Astron. J., 2000, 119: 2498

[55] Lahav O, In: Banday A J, Zaroubi S, & Bartelmann M eds. Mining the Sky, Proc. of the MPA/ESO/MPE Workshop, Germany, Garching, 2000, Heidelberg: Springer-Verlag, 2001: 33

[56] Baccigalupi C, Bedini L, Burigana C, De Zotti G, Farusi A, Maino D, Maris M, Perrotta F, Salerno E, Toffolazzini A, M.N.R.A.S., 2000, 318: 769

[57] Zaroubi S, M.N.R.A.S, 2002, 331: 901

[58] Moretti A, Lazzati D, Campana S, Tagliaferri G, Ap. J., 2002, 570: 502-513

[59] Hoffman Y, In: Banday A J, Zaroubi S, &Bartelmann M eds. Mining the Sky, Proc. of the MPA/ESO/MPE Workshop, Germany, Garching, 2000, Heidelberg: Springer-Verlag, 2001: 223

[60] Chang Tzu-Ching, Refregier A, Ap. J., 2002, 570: 447

[61] Odewahn S C, Cohen S H, Windhorst R A, et al. Ap. J., 2002, 568: 539

[62] Fruchter A S, Hook R N, PASP, 2002, 114: 144

[63] Lasenby A, Barreiro B, Hobson M, In: Banday A J, Zaroubi S, & Bartelmann M eds. Mining the Sky, Proc. of the MPA/ESO/MPE Workshop, Germany, Garching, 2000, Heidelberg: Springer-Verlag, 2001: 15

[64] Hobson M P, Jones A W, Lasenby A N, Bouchet F R, M.N.R.A.S., 1998, 300: 1

[65] Sanz J L, Barreiro R B, Cayon L, Martinez-Gonzalez E, et al. Astron. Astrophys., 1999, 140: 99

[66] Miller C J, Nichol R C, Astron. J., 2001, 555: 38

[67] Szapudi I, et al., Ap. J., 1999, 517: 54

[68] Starck J L, Murtagh F, Querre P, *et al*. Astron. Astrophys., 2001, 368: 730

[69] Kim R S J, Kepner J V, Postman M, *et al*. Ap. J., 2002, 123: 20

[70] Louys M, Starck J L, Bonnarel F, *et al*. Astron. Astrophys. Suppl. Ser. 1999, 136: 579

§1.3.3　离群数据的探索和研究

随着天文数据的急剧增长，这就急需强而有效的分析方法，来充分挖掘隐藏在数据中的信息。数据库中的知识发现(Knowledge Discovery in Databases, KDD)就是从数据中证认出正确的、神奇的、潜在有用的、重要的、可理解的模式的非平凡过程。有关这方面的知识可参考天文学中的数据挖掘和知识发现的详细评述[1]，按照发现的模式的种类的不同，数据库中的知识发现的任务也各不相同，通常分为 4 类[2,3]，① 依赖性探测(如相关规则)；② 种类的证认（如分类和聚类）；③ 种类的描述（如概念的推广）；④ 例外或离群数据（outliers）探测。前三种任务主要针对数据中的大多数数据并从中证认出模式的。大多数数据挖掘的研究属于知识发现任务的前三类。例如：聚类的目的就是发现一组种类或类别来描述整个数据结构。分类的目的是找到一个函数将每一个数据点映射到几种给定的类别中。另外，一个重要的知识发现任务正好与前面的任务相反，其是针对数据中的那些偏离大多数数据分布的数据点，即例外和离群数据（outliers）。寻找例外和离群数据在数据挖掘和知识发现领域并未引起足够重视，仅仅将其作为副产品噪音而被忽略或抛弃。一些机器学习和数据挖掘算法尽管考虑了离群数据的存在，但仅仅在某种程度上容忍它们的存在。但是，在某种情况下，一个人的噪音恰是另一个人的信号。事实上，生活中不凡这样的例子。例如信用卡欺骗的探测、在电子商务中发现罪犯活动、录像监视、药学研究、专业运动员的成绩评估和气象预测等方面发现稀有事件时比通常情况更有趣更有用。在天文学中也不例外，发现稀有的、未知种类的天体和天文现象是天文学家尤其关心和关注的课题。

离群数据的定义

目前还没有一个一般的、统一的、大众普遍接受的离群数据的定义，下面给出几种定义：

⑴ Hawkins 定义（Hawkins-Outlier）[2,4]

离群数据是指其观测值远远偏离其它观测值，以至于怀疑它是由其它机制产生的。

⑵ DB(pct, dmin)-Outlier[4]

在数据集 D 中，如果到离群数据 p 的距离大于到其距离的最小值 dmin 的数据占总数据的至少百分之 pct，则数据 p 称为 DB(pct, dmin)-Outlier。

⑶ 局部离群数据(Local Outlier)[4-6]

离群数据以它邻域的物体为参照，看其离群的可能性，即计算出每个数据点的离群因子，挑选出离群因子最大的几个作为局部离群数据。

离群数据产生的原因及其影响

从离群数据的定义可知，通常离群数据和类别紧密相连：离群数据是那些偏离数据的主要分布的数据；换句话说，离群数据远离或不属于某一类。离群数据具有的特点：

⑴ 数据的预测值是不同寻常的。

⑵ 数据的平均值是不同寻常的。

⑶ 数据不预期地影响参数估计。

如果数据具备这三个特点的其中一个，通常作为离群数据，研究者需要对其引起重视。离群数据产生的原因通常有：

⑴ 数据项本身的错误，如仪器、天气或人为因素的影响造成观测值的错误。

⑵ 有时样本的非均匀性产生离群数据，一些样本的数目远远小于其它样本，导致这些小的样本常被作为离群数据。

⑶ 对未知的数据结构做出不正确的分布假设，从而导致原本不是离群数据的数据成为离群数据。

⑷ 在误差分布的两翼上，极值出现的频率远大于在预期的正常分布上出现的频率，因而这些极值常被当作离群数据。

⑸ 稀有事件或现象的产生。

离群数据的存在，改变了平均值、变量和回归参量，增大了均方差，使估计或预测出现偏差，从而导致错误的结论。另外，离群数据的研究之所以引起我们兴趣的原因还有两个[7]：首先，在数据分析之前，需要平衡各类样本数目。若某些种类样本数偏多，而其它样本数偏少，很可能造成将少的样本当作离群数据去掉，所以应去掉样本数较多的样本中的离群数据以使各类样本均衡。这样的离群数据勿需进一步研究；其次，某些离群数据会给我们新的洞察力，需要对其认真审视和研究。比如在天文学中发现一些稀有的、新类型的天体，这将导致新的理论的发展和完善，高红移类星体、褐矮星的发现即属之列。在大的数据组中通过系统地寻找参数空间中的离群数据可以发现稀有的未知类型天体，SDSS 和 DPOSS 研究组在寻找高红移类星体时就是利用这种方法发现了一些新类型的天体。1998 年 Djorgovski 等[8]在 DPOSS 巡天中利用色参数空间发现了高红移类星体和 II 型类星体。在色参数空间中，正常恒星分布呈现香蕉形状，并形成一个温度序列，而类星体具有不同于正常恒星的颜色，远离恒星区，从而以在色空间的不同位置就可以将类星体与恒星区分开[9]。然后再根据吸收线和发射线的特点就可以区分开高红移类星体和 II 型类星体。这两类类星体相当稀少，面密度小于每平方度 10^{-2}，低于可靠的恒星与星系分类的界限。这样，为了统计性地探测

一些有意义的样本，就需要大天区巡天并且依靠合适的选择方法。同样的方法也适合于其它波段的低角分辨率的巡天，例如：在 IRAS 数据中区分恒星和星系，用射电指数把类星体从射电星系中辨别出来，用 X 射线的硬度比从样本中找出 AGN 等等。在可见光和近红外波段，用同样的方法可以找到各个红移的类星体的完备样本[10-11]。同样，可以选出特殊谱类型的恒星用以探测银河系结构[12]。如果星系的形态可以用适当的方法参数化，则同样也可以将星系区分开[13]。在未探索的参数空间中系统地寻找离群数据可以发现另外一些特殊天体，其中一些结果便是新天文现象的原型。如果这些新的天体或天文现象确实存在，而且可以在已有的数据中探测到，那么彻底的、大范围内无偏差的多波段的宇宙探索将可以发现它们。因此在统计分析、机器学习、模式识别、数据挖掘之前，我们都需要对数据预处理，去掉或挑出离群数据，按其产生的原因来取舍。若是误差或噪音，则去掉；若是稀有或特殊事件，则需要详细研究。

离群数据的探测方法

按变量来分，离群数据分为单变量的离群数据和多变量的离群数据。最简单而普遍的离群数据的研究集中于单变量离群数据的确认[7]。在单变量的情况下，极值显然是离群数据。当数据分布对称时，左右两边的尾部(tails)极值点很可能是离群数据。对应的两边的点分别称为下离群数据(lower outliers)和上离群数据(upper outliers)。数据分布不对称时，分布较宽的一侧的尾部极值可能是离群数据。相比之下，因为多变量的数据分布不存在尾部，使得探测多变量的离群数据比较困难。当然多变量的离群数据有时恰是单变量的离群数据，然而毕竟这种情况为数很少，通常一些在一维上正常分布的数据点在多维数据空间中将成为多变量的离群数据。

一维离群数据的寻找方法：

以一组数据为例，介绍样本的平均值(Mean)、百分点(Percentile)、中值(Median)、模式(Mode)、范围（Range）、变量（Variance）、标准偏差（Standard Deviation）、变化参量（Coefficient of Variation）的定义，以及离群数据的定义及几个评判是否为离群数据的方法，该样本共有 70 个数据，样本以升序排列如表 1.3：

表 1.3　一个具有 70 个数据的样本

425	430	430	435	435	435	435	435	440	440
440	440	440	445	445	445	445	445	450	450
450	450	450	450	450	460	460	460	465	465
465	470	470	472	475	475	475	480	480	480
480	485	490	490	490	500	500	500	500	510
510	515	525	525	525	535	549	550	570	570
575	575	580	590	600	600	600	600	615	615

⑴ 平均值(Mean)

整个样本的平均值：$\bar{x} = \dfrac{\sum_{i=1}^{n} x_i}{n}$

此处 $\bar{x}$ 为平均值，$x_i (i = 1,2,...,n)$ 为观测值，n 代表样本数。

某类样本的平均值：$\mu = \dfrac{\sum_{i=1}^{N} x_i}{N}$

此处 μ 为平均值，$x_i (i = 1,2,...,n)$ 为观测值，N 代表该类的数目。

该例子的平均值为：$\bar{x} = \dfrac{\sum_{i=1}^{n} x_i}{n} = \dfrac{34356}{70} = 490.80$

⑵ 百分点（Percentile）

一个数组的第 p 个百分点是那个值在该数组中至少百分之 p 的数据为小于或等于此值，以及至少百分之（100- p ）的数据大于或等于此值。假如数据按升序排列，计算第 p 个百分点的位置 i:

$$i = \frac{p}{100} \times n ,$$

如果 i 不是整数，则对其取整，第 p 个百分点即是处于第 i 个位置的数据；若 i 是整数，第 p 个百分点则为第 i 个和第 $i+1$ 个数据点的平均值。

该例的第 90 个百分点的位置　　$i=\frac{90}{100}\times 70=63$

其数值为第 63 和 64 个数据的平均值　　$\frac{580+590}{2}=585$

⑶ 中值（Median）

在一个有序的样本中，处于最中间的值，即中值$\overleftrightarrow{x}$。

$\overleftrightarrow{x}=x_m$　　当n为奇数。

$\overleftrightarrow{x}=\frac{x_m+x_{m+1}}{2}$　　当n为偶数。

该例的中值是第 50 个百分点（percentile），

$i=\frac{p}{100}\times n=\frac{50}{100}\times 70=35.5$　　i为数据点的序号，p为此数据点的百分点序号。

$\overleftrightarrow{x}=\frac{475+475}{2}=475$

⑷ 模式（Mode）

样本中出现频率最高的数据为模式。

例如 450 在该样本中出现的次数最多（7 次）。所以其模式为 450。

⑸ 四分点(Quartiles)

四分点是具体的百分点，如第 1 个四分点(Q1)是第 25 百分点，第 2 个四分点(Q2)是第 50 百分点即中值，第 3 个四分点(Q3)是第 75 百分点。则四分点范围(interquartile range, IQR):

$$IQR=Q3-Q1$$

下内限：　　$Q1-1.5\times IQR$

上内限：　　$Q3+1.5\times IQR$

下外限：　　$Q1-3.0\times IQR$

上外限：　　$Q3+3.0\times IQR$

弱离群数据（mild outliers）是处于下内限外和下外限内或上内限外和上外限内的数据。而强离群数据（extreme outliers）则为下外限和上外限外的数据。

在该例中下内限和上内限分别为：

$$Q1-1.5\times IQR=450-1.5\times 75=337.5$$
$$Q3+1.5\times IQR=525+1.5\times 75=637.5$$

所以该例子中不存在离群数据。

⑹ 范围（Range）：

范围（Range）代表最大值与最小值的差值，是一种最简单的弥散检测方法，

从其定义我们很容易看出其极易受到最大值和最小值的影响。

⑺ 变量（Variance）

变量（Variance）代表每一个数值与平均值差值的平方和的平均值。

对整个样本，变量s^2表示为：

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

对某类样本，变量σ^2表示为：

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}$$

⑻ 标准偏差（Standard Deviation）

标准偏差是变量的正的平方根。对整个样本，标准偏差为s；对某类样本，标准偏差则为σ。

⑼ 变化参量（Coefficient of Variation）

变量的参量表示相对于平均值，标准偏差有多大。

对样本而言，变量的参量为

$$\frac{s}{\bar{x}} \times 100$$

对某类样本，表示为

$$\frac{\sigma}{\mu} \times 100$$

在该例中，变量为

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} = 2996.16$$

标准偏差为

$$s = \sqrt{s^2} = \sqrt{2996.16} = 54.74$$

变化参量为

$$\frac{s}{\bar{x}} \times 100 = \frac{54.74}{490.80} \times 100 = 11.15$$

几种确定一维离群数据的数值方法：

⑴ z –Scores

z –Scores 常称作标准值，代表观测值偏离平均值的值与标准偏差的比值，即

$$z_i = \frac{x_i - \bar{x}}{s}$$

本例中最小值的 z –Scores

$$z = \frac{425 - 490.80}{54.74} = -1.20$$

离群数据通常是数据中的非常大或非常小的数据，数据具有大于 3 或小于-3 的 z –Score 的数据被认为是离群数据。离群数据可能由于错误记录或混入其它本不属于该数组中的数据而产生的，亦或是属于本样本中的数据。在本例中，最大和最小的 z –score 是 2.27 和-1.20，利用 $|z| \geq 3$ 作为离群数据的标准，则不存在离群数据。

⑵ Chebyshev's 定理

Chebyshev 定理：数据中至少有 $(1 - \frac{1}{k^2})$ 的数据在其平均值的 k 个标准偏差范围内，此处 k 为大于 1 的整数。例如：

当 $k = 2$ 时，至少有 75%的数据在平均值的 2 个标准偏差范围内。

当 $k = 3$ 时，至少有 89%的数据在平均值的 3 个标准偏差范围内。

当 $k = 4$ 时，至少有 94%的数据在平均值的 4 个标准偏差范围内。

⑶ 经验规则（the Empirical Rule）

对具有钟形（bell-shaped）分布的数据，如图 1.12。大约 68%的数据在平均值的 1 个标准偏差范围内，95%的数据在平均值的 2 个标准偏差范围内，99%的数据在平均值的 3 个标准偏差范围内。

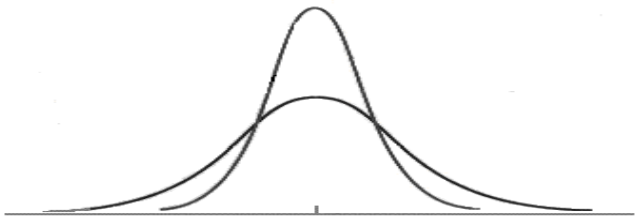


图 1.12　具有钟形分布的数据

为形象直观地发现离群数据，可用散点图（scatter plot）、框图(box plot)、直方图(histogram)，如图 1.13a、b、c 所示。在处理一维离群数据时，Barnett 和 Lewis 针对不同的分布（正态的、泊松的、指数的和二项式的分布）提出约 100 种不一

致检验或离群数据的检验[14]。要选择哪一种检验依赖于：1：数据的分布；2：是否分布参数已知；3：预期的离群数据数目；4：甚至离群数据的种类。但是，所有的检验存在两个严重的问题。首先，这些数据是单变量的。由于这一限制，这些检验不适用于高维变量的情形；其次，其是以数据的分布为基础的，在分布未知的情况下不适用。实际上，我们并不知道数据是属于哪一种分布，正态分布和伽玛分布等，故必须尝试各种检验，以期找到合适的数据分布。

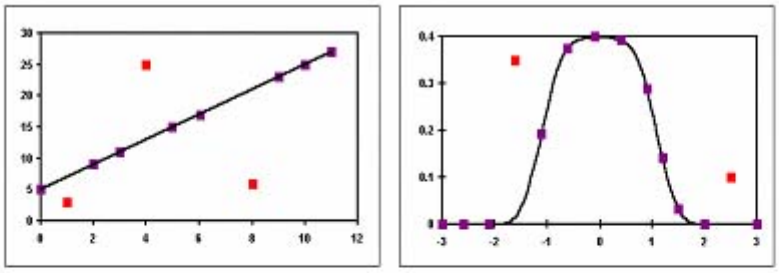


图 1.13a　散点图（scatter plot）

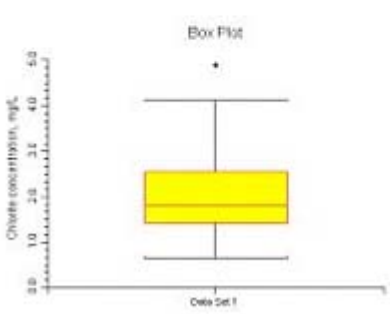


图 1.13b　框图(box plot)

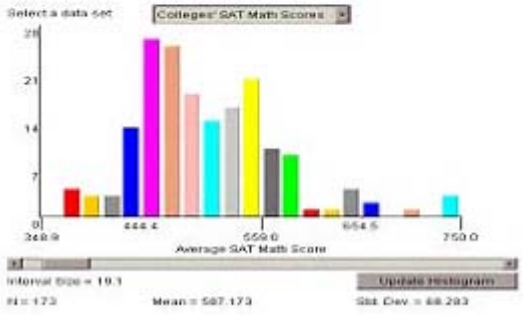


图 1.13c　直方图(histogram)

多维离群数据的寻找方法：

证认一维离群数据的方法是建立在数据的排序上的。但是，对于多变量数据，不存在无争议的排序标准。Barnett 和 Lewis[14,15]提出了一种与众不同的子排序(sub-ordering)方法，简化的子排序方法在离群数据的研究中普遍应用。简化的子排序方法建立需两步[7]。首先，转换每一个多变量观测值 x_i 为标量 r_i，形成一组标量 $R=\{r_i\}(i=1,...,N)$。R 按实际的多变量数据的顺序排序。这种转换是以距离标量为基础进行转换的，故极值是那些具有最大值 R 的多变量观测值。

推广的距离的度量[7,14,15]：

$$r_i^2=(x_i-x_0)'\Gamma^{-1}(x_i-x_0), \qquad (1.2.1)$$

此处 x_0 代表数组的平均矢量，Γ^{-1} 为权重变量与数据的弥散成反比。这些参数的选择不同将导致不同的距离度量。例如 Γ 为单位阵时，（1.2.1）式代表 x_i 到数据中心 x_0 的欧几里得距离。

若选用马氏距离(Mahalanobis distance)[16,17]，只需将（1.2.1）式的Γ变为类的协变矩阵∑即可。通常，类的平均值μ用作中心参量。然而类的平均值常常未知，故常用样本的平均矢量m和样本的协变矩阵S来估计。

$$r_i^2 = (x_i - m)' S^{-1} (x_i - m)，\qquad (1.2.2)$$

马氏距离合并了属性间的依赖关系。这一点在多变量的离群数据探测中尤为重要，因为其目的是要探测不同寻常值的合并的情形。许多距离度量包括欧几里得距离仅利用中心位置（location）信息，因此不适合这种任务。马氏距离的另一个优点是每一个变量标准化为零平均值和单位变量，这样单位变量对距离不会产生影响。

伽玛概率统计分布图在利用推广的距离度量时探测离群数据比较有用。这些图画出了按顺序排列的间约化的单变量r_i随伽玛分布的四分点变化。如果多变量观测值服从正常分布，那么简约的量r_i近似伽玛分布。因此，数据点应围绕一条直线聚类，那些明显偏离线性关系的被认为是离群数据。在不能确定数据分布是否服从正态分布时，我们最好利用框图。因为伽玛概率统计分布需要专家来评估反常数据点是否确实为离群数据，而框图则提供了确定离群数据存在的客观标准。尽管离群数据的不一致检验有坚实的统计理论基础，缺点是假设的每一条必须按步就般地符合。在实际应用中，若一些假设不符合时，利用众所周知的、广泛应用的非正式方法如框图将是比较合适的选择。但是，将来的工作还是应体现出正式的统计检验。我们需注意利用马氏距离探测多变量的离群数据的方法时有两点限制，首先：数据是定量的正态分布；其次，缺值数据在计算距离之前应处理。也许某种不同寻常的距离函数可以解决这个问题[7]。

在计算统计学领域发展了许多适合多维数据的离群数据探测方法。关于离群数据探测的方法林林总总，通常分为 5 类[5,6]。第一：基于分布的(distribution-based)，离群数据是那些偏离正常分布（如正态分布、泊松分布等）的数据。第二：基于深度的(depth-based)，该探测方法依赖于计算 k 维凸球的不同层，离群数据是那些处于球的外层的数据。第三：基于距离的(distance-based)，Knorr 和 Ng 提出的 DB(pct，dmin)-Outlier 方法和探测离群数据的统一方法（a unified approach for mining outliers）。第四：基于聚类的(clustering-based)，一些聚类算法如 CLARANS、DBSCAN 和 BIRCH 可以处理存在离群数据的数据，但是它们的主要目的是聚类，第五：基于密度的(density-based)，Breunig 等[5]提出的局部离群数据（local outliers）发现的方法 OPTICS-OF，Jin Wen 等[6]的挖掘最离群的 n 个局部离群数据(mining top-n local outliers)的算法。还有一种是基于偏差的（deviation-based），Aggarwal 和 Yu 提出的遗传算法[18, 19]。以及其它方法如贝叶斯方法[20]、分形为基础的小波方法[21]、模糊集理论[22]、并行算法[23]和所有的

非监督聚类方法（如自组织映射 Self-organization Map，SOM、主分量分析方法 Principal Component Analysis, PCA）等。

与聚类算法的关系

聚类算法主要是将具有相似属性的事物归为一类，并不是为探测离群数据而设计的[3]。像 CLARANS、DBSCAN 和 BIRCH 算法是专为数据挖掘而设计的聚类算法。在 CLARANS 中，如果某一数据项去掉能提高聚类因子，那么该数据项将被当作噪声去掉。同样在 BIRCH 中，如果某数据相比较偏离离它最近的聚类中心，也会被当作噪声处理。在这两个算法中，离群数据的定义是通过聚类间接定义的，而且这些算法的发展是为优化聚类，而非离群数据探测，离群数据仅作为聚类分析的副产品。在统计学中，发展了若干聚类算法，这些算法通常分为两类：分割(partitioning)算法和分层(hierarchical)算法。在这两种情况下，每一数据项至少属于一类，而且它们目的不是为证认离群数据。在机器学习界的所有聚类算法也如此。区别于 CLARANS 和 BIRCH，DBSCAN 提供了较直接的离群数据确认方法。它通过在 ε 邻域内物体的数目和所考虑的数据的可扩展性（reachability）和连通性(connectivity)，将数据分为核区、边界和离群，设定 ε 足够小以获得较强的聚类。DBSCAN 作为聚类算法主要是想产生最大数目的数据分类，而非标出哪些数据是离群数据。因此，不能简单地将聚类算法拿来用即可，毕竟它们的着重点在聚类而非离群数据。真正充分挖掘离群数据需要发展与之匹配的挖掘算法。

结论

随着天文仪器和观测手段的进一步提高，天文数据以 TB、甚至 PB 计量，数据维数由几维上升到几十维、几百维，这对算法的要求日趋严格，对天文学家也提出了新的挑战。显然，天文学家仅知道本领域的知识是远不够的，需要与统计学家、计算机学家、数据挖掘学家的合作，以应对形势发展的需要。目前，大部分探测离群数据的算法是基于低维空间，适于高维空间的算法还待进一步探讨和研究。而且，每一算法都有其适用范围，作为天文学家要融合数据挖掘、统计学、机器学习和模式识别等学科的优点，探讨出一些切实而有效的适合天文数据特点的挖掘算法。拥有一套适合天文学科的算法，挖掘出隐藏在数据中鲜为人知的、稀奇的、新类型的天体和天文现象，这将推动天文学甚至其它学科的理论的发展和完善。相信在不久的将来，随着国际虚拟天文台（IVO）的创建和运行，天文学家将可以轻松自如地在线数据挖掘。

参 考 文 献

[1] 张彦霞, 赵永恒, 崔辰州, 天文学进展, 2002, 20(4): 312

[2] Knorr E M, Ng R T, Tucakov V, VLDB Journal, 2000, 8(3-4): 237

[3] Knorr E M, Ng R T, Proc CASCON, 1997: 236
http://citeseer.nj.nec.com/232267.html

[4] Breunig M M, Kriegel H -P, Ng R T, Sander J, Proc ACM SIGMOD, 2000: 93,
http://www.dbs.informatik.uni-muenchen.de/Publikationen/Papers/LOF.pdf

[5] Breunig M. M., Kriegel H -P, Ng R T, Sander J, Proc of 3rd European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD'99), 1999: 262,
http://www.dbs.informatik.uni-muenchen.de/Publikationen/Papers/PKDD99-Outlier.pdf

[6] Jin Wen, Tung A K H, Han Jianwei, Proc ACM SIGMOD, 2001: 427,
http://www-faculty.cs.uiuc.edu/~hanj/pdf/kdd01.pdf

[7] Laurikkala J, Juhola M, Kentala E, In: Lavrac N, Miksch S, Kavsek B, eds. The Fifth Workshop on Intelligent Data Analysis in Medicine and Pharmacology, Berlin, 2000: 20

[8] Djorgovski S G, Gal R R, Odewahn S C, et al. In: Colombi S, Mellier Y, Raban B,, eds. Wide Field Surveys in Cosmology, Paris: Editions Frontieres, 1998: 89.

[9] Djorgovski S G, Mahabal A A, Brunner R J, et al. In: Brunner R J, Djorgovski S G, and Szalay A S, eds. Virtual Observatories of the Future, San Francisco: Astronomical Society of the Pacific, ISBN: 1-58381-057-9, 2001: 52

[10] Wolf C, et al. Astron. Astrophys, 1999, 343: 399

[11] Warren S, Hewitt P, Foltz C, M.N.R.A.S, 312: 827

[12] Yanny B, et al. Ap. J., 2000, 540: 825

[13] Odewahn S C, Windhorst R, Driver S, et al. Ap. J., 1996, 472: L13

[14] Barnett V and Lewis T, Outliers in Statistical Data, John Wiley & Sons, New York, 1994

[15] Barnett V, Journal of the Royal Statistical Society A, 1976, 139: 318

[16] Jain A K, Dubes R C, Algorithms for Clustering Data, Prentice Hall, New Jersey, 1988

[17] Boberg J, Cluster Analysis: A Mathematical Approach with Applications to Protein Structure, Academic dissertation, Turku Centre for Computer Science, Turku, Finland, 1999

[18] http://citeseer.nj.nec.com/482998.html

[19] Crawford K D, Wainwright R L, Vasicek D J, Proceedings of the 1995 ACM/SIGAPP Symposium on Applied Computing, Nashville：ACM Press，1995：351

[20] Varbanov A, Technical Report No. 614, June 26, 1996

[21] Struzik Z R, & Siebes A P J M, Technical Report INS-R0008, Febuary 29, 2000

[22] Last M, Kandel A, Proc INTECH, 2001, 2: 292

[23] Hung E, Cheung D W, "Parallel Mining of Outliers in Large Database", in Distributed and Parallel Database, Holand: Kluwer Academic Publishers, Vol 12, 2002, http://citeseer.nj.nec.com/hung99parallel.html

第二章　天体多波段数据的研究和探索

§2.1　活动星系核(AGN)及预选源方法

在离我们极为遥远的诸多星系中，有一小部分星系的核区因其内部有剧烈的物理过程而发出强烈的辐射，成为目前为止我们所知的光度最高的天体。类星体就是这种有着剧烈活动的星系核。在遥远的距离外，这些星系往往只有格外明亮的核区部分才能为人所见，在外观上具有极其类似恒星的点源特性，“类星体”因而得名。因为类星体的光度极高，它们也是我们所能看见的天体中，位于宇宙最深处者。光从遥远的类星体发出，要经过极其漫长的时间才能到达地球，因此我们现在所接收到的来自类星体的辐射携带着宇宙早期的信息。由此类星体可以作为宇宙演化的探针，对大量类星体进行各种统计研究，就能为我们勾画出宇宙从早期到现今的演化图像。然而研究类星体现象必须建立完备的无偏差的类星体样本。类星体数量稀少，位于宇宙最深处。高红移的类星体则更少，搜寻工作非常困难。大量搜寻类星体，构造高质量的完备的类星体样本，这显然是一项难度大、周期长的工作。由于类星体的点源特性，其极易混杂在恒星群中。这样搜寻类星体的难度就在于如何将它们辨别出来。从观测的角度而言，类星体是以其独特的光谱而区别其它天体。而采集光谱需要大量的望远镜的观测时间，考虑到望远镜观测时间的昂贵性，我们最好预选出类星体候选体，以提高望远镜的利用效率。天文学家从来就没有放弃在这方面的探索和努力，并已经取得较大的进展。

下面简要地从活动星系核的特点与分类、预选源方法来介绍一下这方面的工作。

§2.1.1　活动星系核的特点和分类

活动星系核(AGN)的特点：

⑴ 点源特性

这一特征是活动星系核中类星体的最明显的视觉特征。通常由于类星体的流量远大于寄主星系的流量，它们在可见光波段的图像经常表现出一个亮点。许多类星体的 X 射线光度与可见光光度的比值远远大于正常星系的。正是基于这点，它们在 X 射线波段图像也呈现出明显的点源特性。但是，在射电波段则呈现出相当大的展源特性，甚至大于星系。

⑵ 高光度

目前，已知的活动星系核光度覆盖了大约从每秒 10^{42}~10^{48}尔格的能量范围，

而典型的场星系的光度约为每秒 10^{44}尔格。换句话说，活动星系核光度是典型的场星系的光度的 10^{-2}倍到 10^{4}倍。

⑶ 宽波段连续谱

之所以称活动星系核的“宽波段”连续谱，这是相比较于正常星系而言的。作为一级近似，星系光谱无疑是恒星光谱的叠加。恒星光谱的零级近似是黑体谱，故恒星的大部分光度范围较窄，星系的光度范围相应也很窄。大部分活动星系核的谱明显不同于正常星系的谱，其谱是扁平的，覆盖了射电、红外到最硬的 X 射线波段。在射电波段的热辐射高于正常星系的一个量级，甚至几个量极；在 X 射线波段是三到四倍的量级。

⑷ 发射线

活动星系核的发射线备受关注的原因有两个。第一，发射线很突出，等值宽度通常为 100 埃，这明显不同于具有少量的弱发射线而以吸收线为主的恒星和星系。

⑸ 光变

光变也是活动星系核的特征之一。不象正常星系，在可见光波段活动星系核很容易看出改变。在几年的范围内，典型的变化幅度为 10%左右。并且变化幅度似乎有随波长变短而增加的趋势。强的光变特性仅仅与其它三个特性相连：强偏振、致密的射电结构、强的γ射线发射。

⑹ 偏振

大部分恒星就其本身而言是无偏振的，但是光经过恒星际灰尘后会引起约 0.5%的线性偏振，对星系也如此。典型的活动星系核的线性偏振约为 0.5~2%，有点甚至高达 10%。尽管大多数活动星系核存在弱偏振，但这足以与恒星区分开。

⑺ 射电发射

强的射电发射也是活动星系核的一个显著特征。许多已知的活动星系核都是强的射电源。尽管射电波段辐射很强，但还占不到热光度的 1%。而且大量的巡天资料表明绝大部分活动星系核在射电波段的辐射仅为总能量的一小部分。

活动星系核(AGN)分类：

分类，即是根据研究对象的相似性把它们分为几个类群。几乎每一个科学领域都会涉及到分类的问题。科学的发展离不开分类，而且正确的分类又必然导致正确的理论，使人们对事物的认识逐步深化。因而要想揭开活动星系核的面纱，首先需对其分类。通常星系分为正常星系和活动星系；活动星系包括具有活动星系核的星系和其它星系。下面介绍几种活动星系核：

类星体（quasars）：光度最大的活动星系核组成，在可见光波段呈现点源，

且少部分是强射电源。因此根据射电的强弱，又将其分为射电强（Radio-loud）类星体和射电宁静（Radio-quiet）类星体。

赛弗特星系(Seyfert galaxies)：依据与寄主星系的光度对比情况定义赛弗特星系和类星体，如果寄主星系是可见的，则为赛弗特星系，否则为类星体。就拿光度而言，赛弗特星系比类星体弱两个量级，所以赛弗特星系的寄主星系是可见的。

低电离核发射区（LINERS）：象赛弗特星系一样，具有强的发射线，只是来自低电离区的发射线的强度远远超过赛弗特星系的。

蝎虎座 BL 天体(BL Lacs)：其连续谱即无发射线又无吸收线，由于其原型位于蝎虎座且具有光变而得名。

光学激变变星（Optical Violently Variables，OVVs）：在光学波段具有快速而大幅度的光变。因其与蝎虎座 BL 天体有许多共性且是射电源，故二者有时统称为“blazars”。

还有一些类别如宽线射电星系、窄线射电星系、窄线X射线星系等等，这里不再一一介绍。详细的内容可参看有关文献[1,2]。

§2.1.2　预选源方法

活动星系核尤其是类星体由于其大红移及高光度一直成为天体物理研究的热点, 对于研究早期宇宙物质分布, 星系的形成与演化有着重要价值。活动星系核又是比较稀有的天体, 为了寻找它们, 人们花了极大的时间和精力。为了提高效率，人们采取了很多预选源方法，如紫外超、无缝光譜、多色测光及多波段交叉证认等, 这些方法都各有优劣[3]。

⑴ 可见光颜色

通常寻找活动星系核是以它们极宽的连续谱作为依据的。这样在可见光波段，活动星系核与恒星和星系具有不同的颜色。大多数的巡天喜欢用颜色标准来挑选活动星系核而不用其它方法的原因，一方面由于人们擅长测光技术；另一方面则由于这种方法较有效——几个图像就足以挑出大量的候选样本。在实际的巡天中，人们需要拍下某天区至少两个滤光片的图像。在给定天区的图像中，大部分天体并不是活动星系核。若将活动星系核证认出来，要首先预选源。颜色选源方法的优点：通过活动星系核具有不同于恒星和星系的可见光谱形状和不同波段的流量比（即颜色的不同），很容易选到活动星系核。它们的区别来源于若干方面：第一：活动星系核的连续谱宽而平滑，而恒星和星系的谱为热谱且有较强的弯曲。典型的活动星系核的谱宽度意味着相对于 V 波段，它们在紫外和红外波

段的流量远大于恒星。第二：当红移足够大使得赖曼α线移至可见光波段，强的发射线必然加大了它移至的波段的流量。第三：当红移足够大时，介于赖曼α线和赖曼连续谱之间的吸收线，则会降低它移至的波段的流量。

当然如果活动星系核与其寄主星系的颜色明显不同于正常星系的颜色，这说明活动星系核与其寄主星系的光度相当或更大。因此颜色的选择效应会自动偏离那些低光度的活动星系核，同时会带来另一个不明显的偏差。若选取附近的活动星系核，则希望源为展源以排除掉恒星；若选取较远的活动星系核，则希望为点源以去除掉星系。前一标准易错过那些比其寄主星系亮的活动星系核，后一标准则相反。

颜色选源的具体做法：先考虑在一维色空间中 U-B 色指数的截断。最热的主序恒星的 U-B≈0.4-0.5mag，相比而言，典型的活动星系核谱的 U-B≈-1mag。若以紫外超（UV-excess）为标准，大部分恒星和星系自然排除掉。主要的污染物是那些热的亚矮星、白矮星，只要降低流量极限，去掉这些污染物不再是问题。当然由于恒星在空间中的分布不同，低银纬聚集，高银纬离散，在高银纬恒星的污染则会减少，而低银纬（尤其银纬 10 度以内）寻找活动星系核则较困难。

最早期的开创性的色巡天工作之一应追溯到 1967 年Markarian的工作。在Markarian表中约有 1500 个星系，约 10%为赛弗特星系。最详细的颜色巡天工作是 1983 年Schmidt和Green的帕洛玛亮的类星体巡天(Palomar Bright Quasar Survey，BQS)。他们在 11000 平方度的天区仅选取U-B<-0.44 mag的天体，它们发现了约 1700 个亮于m_B≈16 mag。通过光谱证认，他们确认了 114 个天体为活动星系核，正确率约为 **7%**。

为了克服仅利用紫外超（UV-excess）选源的不足，人们开始用多色方法来选源。例如：Warren等人[4]拍下五个波段的图像。发现恒星在四维色空间中分布似“香肠（Sausage）”。因此他们选取那些具有类点源图像的在任何方向偏离恒星分布的“香肠”区的天体为活动星系核的候选体。这样他们发现了许多用单纯的紫外超方法不能发现的活动星系核。这种多色方法成为SDSS(the Sloan Digital Sky Survey)巡天选源的方法之一。选取 10^6个点源的类星体候选体，约 10^5个（大部分是最亮的，占到~**10%**）通过光谱观测得到证认。

但是，无论是紫外超方法还是多色方法应用于可见光巡天都会遇到不可避免的困难：在我们观测者与活动星系核之间存在着灰尘或其它物质的吸收。这些吸收会随着活动星系核的红移的增加而影响活动星系核的表现特征，从而增加发现活动星系核的难度。

⑵ 可见光发射线

活动星系核的另一个突出特征是它们具有强的可见光和紫外波段的发射线。

这些线即使在粗糙的低色散的光谱中也较容易发现。同时这也是区别于恒星的重要特征，因为几乎没有恒星具有这样的发射线。所以可以利用这一特征来预选活动星系核。利用拍到的所有活动星系核候选体的高质量的光谱，可以证认活动星系核，并且可以准确地测量它们的一些性质如红移、发射线流量和连续谱的流量。由于在静止坐标中紫外发射线的特征较明显，因此这种方法对发现高红移类星体尤为有用。历史上人们是通过人眼来识别光谱，这不可避免不同的人有不同的选择标准。后来为避免人为性和非普适性，人们将这种方法自动化。例如：为建立大而亮的类星体样本，天文学家利用极端蓝的颜色、强的发射线或吸收线特征、或强的连续谱截断，发现了约 1000 个类星体。Schmidt，Schneider 和 Gunn 利用发射线的等值宽度和信噪比挑选活动星系核候选体。在一些以发现恒星和星系为主的巡天（如 CfA、SDSS），仍利用光谱特征，一些活动星系核是作为副产品被发现的。

⑶ X 射线

许多活动星系核的X射线与可见光波段流量的比值达到 1，而恒星和星系却达不到。即活动星系核是强的X射线源，因而X射线巡天又是一个有效的寻找类星体的方法。HASS巡天(HEAO-1 All-Sky Survey)和EMSS (the Einstein Medium Sensitivity Survey)与ROSAT深视场巡天（ROSAT deep survey）是获得活动星系核在X射线波段统计信息的主要源泉。前者对应硬X射线波段 2-10KeV；后者对应软X射线波段 0.1-2.4KeV。通常硬X射线巡天覆盖大的天区而较亮的流量极限；软X射线巡天覆盖小面积深视场。例如EMSS巡天仅覆盖了 780 平方度而流量极限小到 6×10^{-14}erg $cm^{-2}s^{-1}$。大约在发现的 1400 个源中 **30%**为活动星系核，而且大部分是低红移的类星体。这说明X射线巡天选择活动星系核的高效性，而BQS巡天利用颜色方法选择活动星系核候选体的成功率仅达 7%。ROSAT深视场巡天的设计方案类似于EMSS巡天，对类似的质子能量范围敏感，但是其能量极限达到 3×10^{-15} erg $cm^{-2}s^{-1}$。该巡天发现了 107 个类星体。另外Piccinotti巡天尽管覆盖的天区较大但仅限于亮源。在其发现的 85 个源的样本中，60 个为河外天体，其中星系团和活动星系核约各占一半。而且所有的活动星系核为I型塞弗特星系。较大的活动星系核的X射线巡天是RASS巡天(the ROSAT All-Sky Survey)以及ROSAT定点观测。总之，这些巡天发现的活动星系核数目约达 10^5个。

⑷ 红外波段

我们也可以在红外波段对比活动星系核的宽的连续谱与星系的窄的吸收谱来发现活动星系核。但是与 X 射线波段相比，寻找活动星系核的效率则偏低，这主要是由于正常星系的一部分辐射在红外波段。另外相比于别的巡天技巧，红外波段还比较不成熟。目前大家比较熟悉的红外巡天是 IRAS 卫星，其角分辨率

为几个角分，极限灵敏度仅约为 0.3-3Jy。不过 IRAS 卫星覆盖了整个天空并且发现了许多重要的天体。类似 CfA 巡天，IRAS 卫星的目的不仅仅为寻找活动星系核。但这里我们只讨论这方面的问题。选择那些热的红外颜色，很容易找到活动星系核。大多数的星系的红外流量在 60-100μ，活动星系核在 12-25μ。换一种说法，正常星系的尘埃的色温度约为 30K，活动星系核通常约为 100-300K。

⑸ 射电波段

历史上第一批类星体是通过射电源与可见光波段的恒星状源的位置巧合发现的。这种方法的主要缺点：对于无偏差的证认要求射电位置的精确度很高约为 1 角分；这只能挑出那些仅占活动星系核一小部分（5-10%）的射电强的类星体，故无法选出活动星系核的完备样本。至于部分活动星系核的射电流量可参考 3C 星表。但是在FBQS巡天中Gregg等人[5]和White等人[6]用射电选源的方法，成功率仅可以达到约 **65%**。FBQS巡天在射电波段的灵敏度足够大，可以发现那些常被当作射电宁静的类星体。射电选源的方法也有自己的好处：它可以发现那些易被可见光选择方法忽略掉的具有不同寻常颜色的类星体。

⑹ γ射线

最后一个发展的巡天计划是高能γ射线巡天。寻找γ射线源在技术上还存在困难，一是难于建立将低流量的信号从噪音中提取出来的探测器；二是达到约 1 度的角分辨率也绝非易事。不过一旦发现，必定很有意义。因为平常的恒星根本不存在γ射线辐射。比如进行γ射线巡天的康普敦γ射线天文台的EGRET望远镜(the Energetic Gamma-Ray Experiment Telescope)，其覆盖全天，能够探测到在30MeV-3GeV能量范围内流量最低达到约 10^{-11} erg $cm^{-2}s^{-1}$的活动星系核。但其获得的位置精度比较粗糙：在 100 MeV质子能量方向的位置精度仅约为 5 度。不过，该巡天发现了若干活动星系核。例如：在 129 个点源中，40 被确认为活动星系核，另外 11 可能为活动星系核，正确率达 **31%**。

⑺ 无缝光谱

无缝光谱的基本思想：几乎同时在一天区获得大量天体的光谱，以此来提高巡天的效率。例如在宽视场的望远镜前面加一光栅。类星体的连续谱和强的发射线明显不同于恒星的，即使在低光谱分辨率也如此，这样较易选取类星体候选体。当紫外超色选择方法选择类星体候选体的效率在红移 z≥2 突然下降时，无缝光谱则对选择高红移类星体尤其有效，这主要是由于强的赖曼α线和 CIV 发射线红移到可见光波段，地面望远镜很容易观测到它们。比较于色选择方法，无缝光谱的类星体的探测效率随红移的改变变化较慢。这一方法的主要缺点是受到无缝光谱的流量极限和有限的波长范围等的限制，及其光谱难以确定巡天的极限星等。

⑻ 光变

类星体在大约一年内典型的光度变化为十分之几个星等。这也正是其有别于大多数恒星的特征。许多工作正是以光变这一特征作为选取候选体的方法。如果有足够多的样本时间跨度好几年，那么用光变来选取类星体的候选体是十分有效的。由于时间的膨胀效应，此方法在选取候选体时会产生（1+z）的统计偏差。光度与光变幅度的逆相关可以导致选取高光度源的偏差。理论上，这些效应可以模拟出并加以校正。

⑼ 零自行

当然发现活动星系核候选体还可以利用别的性质，零自行（zero proper motion）等。利用这一性质可以将类星体从银河系天体（尤其白矮星）中分离出来。由于宇宙学距离，类星体用目前的观测技巧无法探测其自行。这样以零自行为标准可以选择到类星体的完备样本候选体。但是如果样本中含有大量的具有小自行的暗弱星系和银河系天体，这时这种方法就不再有效。必须与别的方法结合方显出其用途。

⑽ 混合方法和其它方法

比较上述方法，我们发现每种方法各有优劣。为减小偏差，最大化类星体候选体数目，人们通常采用几种方法混合使用，优势互补。例如LBQS巡天(the Large Bright Quasar Survey)就采用了色选择方法和无缝光谱两种方法。易被色选择方法忽略的高红移类星体可以用无缝光谱方法选出来，而易被无缝光谱忽略的弱发射线的类星体可以通过色选择方法得到。这样获得的样本比用单一的方法要完备些。Brunzendorf和Meusinger[7]提供了一种不直接依赖类星体的能谱分布的寻找类星体的方法（Variable and proper motion search，间称VPM search ），即选取光变且零自行的天体为候选体。这种方法与用色选择、红移、谱指数、或发射线的等值宽度等方法相比，不具有选择偏差。他们认为这种方法选取候选体的完备性大约 90%，正确率约为 **40%**。另外Meusinger和Brunzendorf[8]指出为增加样本的完备性，最好与色选择方法结合使用。

魏建彦等人[9]的图 1 统计表明 92%的活动星系核处于logC≥0.4R+4.9 的区域（这里的C表示X射线计数率，R为R星等），而O-M光谱型的恒星在该区仅占 3%。因而他们以logC≥0.4R+4.9 为标准选取活动星系核候选体，可以排除掉大多数的O-M光谱型的恒星。他们的具体选择标准：① 未知源在SIMBAD、NED和其它可获得的星表中无对应体；② 赤经 δ ≥3 度；③ 银纬|b|≥20 度；④ 可见光对应体应在以$r=r_1+5''$为半径的圆内，其中r_1为RASS源的位置误差；⑤ 可见光对应体的R星等范围在 13.5 到 16.5 之间；⑥ logC≥0.4R+4.9。他们选取了 165 个未证认的X射线源进行光谱证认，发现 115 个具有发射线的活动星系核、2 个蝎虎

座BL天体、4 个蝎虎座BL天体候选体、22 个星系团、12 个恒星和 10 个仍未证认的天体。用这一标准选择活动星系核候选体的成功率达 **73%**。

为获得较完备的样本，何香涛等人[10]采用多波段的类星体巡天（the Multiwavelength quasar survey, MWQS）方法。在X射线波段，他们选取X射线流量每秒大于 0.02 计数，这样排除掉一些弱源，即一些具有较大不确定流量的很可能是假源的源；接着排除掉那些明显不是活动星系核的源。因为通常白矮星的HR1≤-0.5，恒星的HR1 分布在-0.5 到 0.5 之间，因而选取HR1≥-0.5 的源为候选体。从他们的图 1（横坐标为视星等，纵坐标为X射线流量与可见光流量的比值$\log(f_x/f_o)$）可以看出，恒星具有亮的可见光星等和弱的X射线发射；星系具有弱的可见光星等和中等强度的X射线发射；活动星系核具有最弱的可见光星等和最强的X射线发射。选取亮于 16.0 星等和X射线流量与可见光流量的比值$\log(f_x/f_o)$ ≥-0.5 的源为最后的候选体。在可见光波段，三个主要的选择标准为：① 相对于普通恒星极为蓝的候选体；② 具有强的发射线和吸收线的天体；③ 具有强的连续谱截断。多波段的类星体巡天与大而亮的类星体巡天（the Large Bright Quasar Survey, LBQS）仅有的不同是星等范围，前者为 16.0≤B≤19.0，后者为 16.0≤B≤18.5。因而前者比后者能够探测到更多的类星体。其中三个视场在射电波段的选择标准：① 源的射电位置与可见光位置应在 1.2 角秒内交叉证认；② 可见光的形态至少在两个帕洛玛底片之一中呈现类点源特性；③帕洛玛底片的颜色应蓝于O-E=2。另一个视场在该波段的选择标准：除了满足上述的三个标准外，由于NVSS源的位置精度不高，应在 5 角秒内与APM源交叉证认。按照上述标准，他们选取了 30 个候选体，其中 3 个是已知的类星体，7 个是已在北京天文台的望远镜证认的活动星系核。光谱观测 20 个天体，证认了 12 个活动星系核、4 个恒星和 4 个未确认的天体。因而他们的选源标准的成功率为 **73.3%**（22/30）。

陈阳等人[11]在何香涛等人[10]的工作基础上，修改了他们在X射线波段的选源标准，新的标准为：① X射线流量每秒大于 0.02 计数；② X射线流量与可见光流量的比值$\log(f_x/f_o)$≥-1.0；③ 源在X射线的扩展参数ext满足log(ext)<1.8。这是因为RASS亮源表与各种星表（包括类星体、恒星、星系和星系团表）的相关性表明：当选取可见光星等大于 16 时，将活动星系核从非活动的星系核中挑出来的合理的X射线流量与可见光流量的比值为$\log(f_x/f_o)$ ≥-1.0。既然X射线流量极限自然地将X射线流量与可见光流量的比值定得远大于-1.0，所以$\log(f_x/f_o)$≥-1.0 的标准仅对亮源有效。同时将X射线的扩展参数ext的临界值定为log(ext)<1.8，这样可以包括大多数的活动星系核，排除掉大多数的远距离星系团。利用此三个选择标准得到 98 个X射线候选体。为了观测的有效性，他们又选取可见光星等B≤

19.0，这样剩余 66 个，其中有 12 个是已知的活动星系核。光谱证认了 23 个候选体，发现 9 个活动星系核、8 个恒星和 6 个未证认的天体。从而得到 21 个活动星系核样本。他们的选源标准的成功率为 **31.8%**（21/66）。

BATC计划是利用北京天文台 60/90 厘米施密特望远镜 2048x2048 CCD ，开展的大视场(58x58 角分)多颜色(15 个中带滤光片, 波长覆盖范围 3000 埃－10000 埃)的巡天计划。由于其颜色较多, 15 个颜色的星等组成的SED(Spectral Energy Distribution)相当于一个低色散光谱, 这样BATC系统比其他的宽带多色测光系统有更深的极限星等（天空背景暗）和更高的分光测量精度。使得BATC系统更适合于挑选类星体[12,13]。为充分发挥BATC测光系统的优越性，他们在不同时期采用了不同的方法：模拟方法、双色图方法、类星体模板相关的方法。樊晓晖和陈建生等人对类星体的光谱特性进行了详细的研究。利用模拟方法来研究类星体的各种发射线随着在可见光区加入不同的红移对多色测光色指数的影响。通过研究，他们发现随着红移的增加，位于Ly α 发射线短波方向的各种吸收特征进入了可见光区，内禀的幂率谱和发射线强度分布对类星体色指数及其弥散的影响将是次要的，各种吸收系统的作用将改变类星体的色指数随红移变化的趋势，其中Lyman线系吸收系统的影响最大。樊晓晖和陈建生还利用已知的对类星体的统计结果对高红移类星体光谱进行了Monte Carlo模拟，得到了UBVRI五色测光的色指数，并利用模拟的选择判据计算了选择效应函数。在此基础上，研究了不同红移类星体在BATC测光系统中的表现，确认了BATC测光系统在类星体选择上的优势。樊晓晖将该方法成功地应用于SLOAN巡天数据中，发现了大量的类星体和高红移星系。严皓景将双色图应用在BATC的T329 和T359 两个天区的实测数据中。通过对几种双色图的对比，挑选出在各双色图上都处于天体分布密度较低区域的天体，将这些天体作为类星体的初选结果。排除掉测量误差的因素，再具体分析所有颜色过程的光谱能量分布，挑出许多类星体的候选体。通过观测，证实了BATC发现的第一批类星体。在“双色图”上，找孤立点（离群数据）是挑选类星体的典型方法。一颗星在某两个波段的星等相减值称为它的色指数，若对许多天体在两个以上的波段内都测得了星等值，以某个色指数为横轴，另一色指数为纵轴，所得图即为双色图。双色图直观地显示了天体在所考察波段内辐射流量的比值关系。同一视场中绝大部分天体是恒星，它们的这种比值关系比较固定，因此在双色图上它们密集地分布在一小条较窄的区域。而类星体的辐射方法远不同于恒星，若考察的波段合适，它们在双色图上的位置会远离密集的恒星区而成为孤立点。这些孤立点并不都是类星体，但可以进一步挑出类星体候选体，工作量已大大降低。尽管双色图对选取类星体很有效，但它具有较强的人为因素，且不适合一次性发挥BATC多色的特点。他们发展了一套构造类星体模板，然后与

观测数据相关的方法搜寻类星体。他们通过与模板做线性相关的方法来初步判定它的类型（如恒星可给出光谱光度型，类星体给出红移大小）。他们先将Gunn & Stryker[14]恒星谱与BATC的 15 个滤光片透过滤曲线进行卷积，得出恒星模板SED。将此SED与测光SED作线性相关，给出相关系数。由于对较暗弱天体，其在某些波段的星等值已小于极限星等，会造成某些颜色的缺失。这样不同天体就会有不同数目的颜色，若直接由相关系数判断是否为恒星，就存在因为颜色数不同造成的不等权效应。所以他们由颜色数与相关系数结合给出一个置信度，并给定一个下限。大于此下限则为恒星，并给出摸板中与天体最接近的恒星的光谱光度型；小于此下限的作为反常天体的候选体。鉴于类星体谱的基本特征相同（如强发射线、幂率连续谱等），他们采用类星体合成谱作类星体的模板。首先用三组类星体的合成谱[15-17]合成新的合成谱, 然后将此合成谱从 0 红移至 5.5 红移，红移间隔为 0.01，在每一个红移上与BATC的滤光片透过滤曲线卷积，这样得到了红移 0 至 5.5 的类星体模板。将此模板SED与反常天体的候选者SED做线性相关，与找恒星的步骤类似，找出类星体候选者并给出可能的红移值。这样共从有效颜色大于三个的源中挑出类星体候选体。以上两步由程序自动实现，因此为保险起见，最后将类星体候选体用目视一一确认下来，找出较可能的候选体。张昊彤根据类星体的观测光谱特性和关于类星体方面的理论，建立了适合BATC巡天的搜寻类星体的不同红移的类星体模板。类星体的模板包含了类星体光谱中的三部分变化规律：① 幂率连续谱: $F(\nu)\propto\nu^{\alpha}$, α取-0.25, -0.75, -1.25；② 发射线: 按高斯轮廓模拟发射线，对宽线取FWHM=5000km/s；对窄线取FWHM=1000km/s；各发射线强度比取自Wilkes的文章[18]; 固定线比, 分别取Lyα等值宽度为 65±34Å；③ 根据Madau的文章[19]计算出不同红移处Lyman系限系统及Lyman线系 1-19 根谱线吸收的平均光深。将以上三部分合成即得类星体的人工合成谱。

总之，目前选取候选体的方法的成功率约为 40%-80%。比较上述方法，我们发现这些方法各有优缺点。优点是这些方法方便快捷、选取各参数的临界值或截断点的物理意义明确；缺点是仅限于小样本的研究、截断点的人为性强。随着望远镜和探测器的技术的进步、计算能力的指数增长和收集观测数据的技术的提高，观测天文学正在经历着一场“革命性的变化”。来自空间观测站和地面望远镜的大规模的数字巡天使得天文数据在数据量、质量和复杂性上持续增长。天文数据涵盖了各个波段从射电波段到X射线，甚至 γ 射线波段，总量将达TB，甚至PB，因而Szalay和Gray[20]认为天文学正在面临着“数据雪崩”。面对如此海量的数据，该如何存储、整理、查询、组织和挖掘，这是摆在天文学家面前的难题，同时其也是虚拟天文台的主要任务。典型的数组含有约 10^8-10^9个源，具有 100 个测量属性，也就是说 10^9数据矢量分布在 100 维空间中。要想处理这样的数据，

利用专家知识和一些非自动化的方法显然已经力不从心，因而需要利用现在正在崛起的数据挖掘和数据库中的知识发现的知识来发展适合天文学科的自动化方法。

参 考 文 献

[1] Krolik J H，1999, *Active Galactic Nuclei: From the Central Black Hole to the Galactic Environment*, (Princeton University Press: Princeton)

[2] Peterson B M，1997， *An Introduction to Active Galactic Nuclei*, (Cambrige University Press: Cambrige)

[3] Weedman D W, 1986, Quasar Astronomy(Cambridge University Press),19

[4] Warren S J, Hewett P C, & Osmer P S, 1994, ApJ, 421, 412

[5] Gregg M D, Becker R H, White R L, et al. 1996, AJ, 112, 407

[6] White R L, et al. 2000，ApJS, 126, 133

[7] Brunzendorf J, Meusinger H, 2001, A&A, 373, 38

[8] Meusinger H, Brunzendorf J, 2001, A&A, 374, 878

[9] Wei J Y, Xu D W, Dong X Y, et al. AASS, 139, 575

[10] He X-T, Wu J-H, Yuan Q-R, et al. 2001, AJ, 121, 1863

[11] Cheng Y, He X-T, Wu J-H, et al. 2002, AJ, 123, 578

[12] Fan X H, Burstein D, Chen J S, et al., 1996, AJ, 12, 628

[13] Zhang M, Chen J S, et al. 1998, Progress in Natural Science, 8, 234

[14] Gunn J E, Stryker L L, 1983, ApJS, 53, 121

[15] Zheng Wei, et al. 1997, ApJ, 475, 469

[16] Cristiani S, & Vio R, 1990, A&A, 227, 385

[17] Francis P J, Hewett P C, Foltz C B, Chaffee F H, Waymann R J, & Morris S L, 1991, ApJ, 303, 336

[18] Wilkes B J, 1986, MNRAS, 218, 331

[19] Madau P, 1995, ApJ, 441, 18

[20] Szalay A, Gray J, Science, 2001, 293，203

§2.2　多波段交叉证认

半个多世纪以来，随着科学技术的突飞猛进，天文学正在步入新的阶段。射电技术、红外探测技术、航天和空间技术的兴起和发展，使早先仅限于光学波段的天文观测，一跃成为包括射电、红外、可见光、紫外、X 射线、直至 γ 射线的全波段观测，极大地扩展了人们的视野。来自各个波段的巡天和观测数据急剧增长，如何将这些星表统一起来以探测和研究天体在各波段的特性，这就需要星表之间的交叉证认。证认的可靠性主要依赖于各个波段可对比参数的数目，以及这些参数的质量和准确度。证认来自不同波段的源的纽带是源的位置，因为各个波段的星表的共同参数只有位置。尽管探测器技术不断地提高，位置的准确度仍随波段的不同而不同，这给交叉证认的准确性带来一定的难度。当然其它参数对于评价交叉证认的可靠性也是十分重要的。

交叉证认的原理：

设在两个星表中对应源（即T点和C点）的坐标分别为（α_1，δ_1）、（α_2，δ_2），求它们之间的角距离d（如图 2.1 所示）。分两种情况求解：

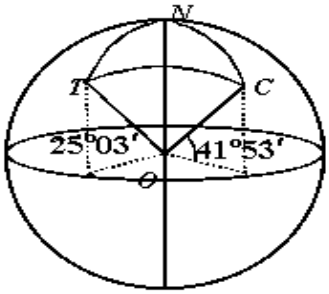


图 2.1　两个源 T 点和 C 点的空间分布

⑴ 若$|\alpha_1-\alpha_2|\leqslant 180$ 度

$\angle TNC = |\alpha_1-\alpha_2|$

⑵ 若$|\alpha_1-\alpha_2| > 180$ 度

$\angle TNC = 360-|\alpha_1-\alpha_2|$

利用球面上的余弦定理：

$$\cos\angle TOC = \cos\angle TON\cos\angle CON+\sin\angle TON\sin\angle CON\cos\angle TNC$$

$$= \sin\delta_1\sin\delta_2+\cos\delta_1\cos\delta_2\cos(\alpha_1-\alpha_2)$$

$$d=\angle TOC=\arccos(\sin\delta_1\sin\delta_2+\cos\delta_1\cos\delta_2\cos(\alpha_1-\alpha_2))$$

通常情况下，在 d 很小时角距离可取如下近似公式：

$$\delta = (\delta_1+\delta_2)/2$$

$$d^2 = ((\alpha_1 - \alpha_2)\cos\delta)^2 + (\delta_1 - \delta_2)^2$$

设两个星表的误差半径分别为r_1和r_2，通常角距离应满足下列条件：

$$d \le |r_1| + |r_2| \qquad (2.2.1)$$

或者：

$$d \le \sqrt{r_1^2 + r_2^2} \qquad (2.2.2)$$

即交叉证认的半径应满足(2.2.1)式或(2.2.2)式。

通过虚拟天文台将天文数据从物理上融合在一起，事实上是一个提取来自不同源泉的多波段星表的相关信息的课题。表面上看起来简单，而实际上该问题受到几个问题的制约：首先，纯数据量的大小问题。通过静态和动态的方式交叉证认来自成千上万平方度天区的多波段数据很显然在计算上是个挑战，即使面对静态的数据库。由于观测数据一直受到相关技术的限制，例如仪器的灵敏度和分辨率随波长的不同而不同，通常可见光到红外波段的分辨率要优于高能波段的。其次，数据的记录质量（光谱、时间或空间上）。数据的质量变动很大，故无歧义地匹配不同波段的源是十分困难的。最后，天空在不同波段的表现不同，使得联合多波段的数据时，会产生一对一、一对多、多对一、多对多、甚至一对无和多对无的源。因此有些时候不同波段源的相关性必须运用概率的方法来确定。

交叉证认的概率

为了充分地研究各种源的性质及宇宙的物理机制，交叉证认来自各个波段的数据显得越来越重要。但是，由于探测器的有限的探测能力和望远镜的有限的分辨本领，所有的观测都具有一定程度的误差。一个源的探测误差半径从VLA的源的几个角秒[1]到EGRET源的一度[2]。因此，证认一个波段的源与另一个波段的源必然存在歧义。这就需要发展一个客观的方法来选择一个源的最可能的可见光对应体，并能给出每个源的证认的可靠性的估计。

天文学家从来未停止过在这方面的探索。Richter[3]首先提出了两个普遍接受的仅匹配高于底片极限的最近源的方法和概率比率的方法，而后由de Ruiter, Willis和Arp[4], Prestage和Peacock[5]，Wolstencrof等人[6]得以修改和完善。利用贝叶斯理论，de Ruiter, Willis和Arp[4]提出了一种定量地估计 1.4GHz射电源的可见光证认的准确率的方法，同时也给出了一种方法用以计算一个证认的射电源样本的可靠性和完备性。用同样的方法，Willis等人[7]和Prestage等人[5]计算射电源的可见光证认的概率；Wolstencroft等人[6]给出了红外IRAS点源的证认概率；Wu等人[8]估计了一个未证认的EGRET源的可能对应体的概率；Mattox等人[9]计算出EGRET源的射电证认概率；Masci等人[10]给出一个新的完备的射电样本的可见光证认概率。

下面详细介绍de Ruiter, Willis和Arp[4]提出的计算可见光证认的概率方法，及如何确定一个证认的射电源样本的可靠性和完备性。

事实上射电源与证认体之间的位置误差（Δα和Δδ），通常射电源的坐标减去可见光源的坐标并不能直接用于确定某个可见光源为对应体，这个复杂性主要由两个因素决定：

① 射电源存在位置误差，赤经的误差介于0.5到5角秒之间。而且赤纬的误差要比赤经大一个因子cosecδ, 因而天文学家估计Δδ比Δα有更大的弥散，事实也确实如此。

② 尽管底片上的可见光天体具有相同的质量，但是天体的密度随纬度的不同而不同。他们的样本分布于银纬|b|≈10 度到|b|≈90 度之间，因此亮源的密度的变化不可忽略。

他们通过源计数的方法来考察密度随银纬的变化情况，拟合 Wr 2C 数据得到了密度分布函数：

$$\rho(b) = (1.40 + 2.39\cos ec|b|) \times 10^{-4} arc\sec^{-2} \qquad (2.2.3)$$

密度在$3.8\times10^{-4} arc\sec^{-2}$（|b|=90 度）与$5\times10^{-4} arc\sec^{-2}$（|b|≈40 度）之间变化。在低于银纬|b|≈30 度时，底片上天体的密度急剧增加，在|b|≈10 度时达到$13.5\times10^{-4} arc\sec^{-2}$。在|b|=12 度的 NGC 6946 星系周围，天体分布极为密集，密度竟达到$21.5\times10^{-4} arc\sec^{-2}$。需要注意在任何银纬，密度的方均根不确定性约为15%-20%。

很显然，证认的标准不仅要考虑射电源的位置误差的变化，而且还要考虑可见光源的密度分布函数。假设射电与可见光的位置的不确定性及可见光源在底片上的密度分布已知，这样确定证认概率可以通过对比如下两个概率：① 考虑到射电源与可见光证认源本质上应处于同一位置和射电与可见光的位置的不确定性，分布在r到$r+dr$的先验概率为$dp(r\,|\,id)$；② 考虑到可见光证认源是混淆的背景源，以射电源位置为中心，在半径r到$r+dr$的区域内，可见光证认源的先验概率为$dp(r\,|\,c)$。如果前一个概率比后一个概率大L倍，我们则认为可见光源与射电源相关。关于这个概率的数学表达式由下面的推导得到。

首先定义无量纲量r：

$$r = \left(\frac{\Delta\alpha^2}{\sigma_\alpha^2} + \frac{\Delta\delta^2}{\sigma_\delta^2}\right)^{1/2} \qquad (2.2.4)$$

这里的$\Delta\alpha$与$\Delta\delta$分别为射电源与可见光源的赤经与赤纬之差（某种意义下，射电位置减去可见光位置），$\sigma_\alpha^2 = \sigma_{\alpha_{rad}}^2 + \sigma_o^2$，$\sigma_\delta^2 = \sigma_{\delta_{rad}}^2 + \sigma_o^2$，$\sigma_{\alpha_{rad}}$与$\sigma_{\delta_{rad}}$分别为射电源的赤经与赤纬的标准偏差，$\sigma_o$为可见光位置的测量误差。通常不管可见光的图像的形状，取$\sigma_o = 1$。

假设最多有一个可见光源离真正的证认体足够近，忽略混淆源比真源离射电源近的可能性，这种可能性的概率为百分之一。现在计算在半径 r 到 $r+dr$ 的范围内找到第一个混淆源的概率。既然背景源的位置分布遵循泊松分布，该概率由下式求解：

$$dp(r\,|\,c)=2\lambda r\times e^{-\lambda r^2}dr \qquad (2.2.5)$$

这里的 $\lambda=\pi\sigma_\alpha\sigma_\delta\rho(b)$ 。(2.2.5)式的表示形式得益于无量纲量 r 的选择。

由于测量误差，射电源与可见光对应体的位置差的概率密度分布遵从瑞利分布(the Rayleigh distribution)。因此在半径 r 到 $r+dr$ 的范围内找到射电源的真实的可见光对应体的概率为：

$$dp(r\,|\,id)=r\times e^{-r^2/2}dr \qquad (2.2.6)$$

这样问题变为区别处于(2.2.6)式的分布的尾部的证认体与混淆的背景源。必须找到一个最优的截断标准，使得放掉那些远离雷利分布的尾部的证认体的几率尽可能地低，但出现第一个混淆源的几率会尽可能地高。因此，他们构造了概率比率 LR：

$$LR(r)=dp(r\,|\,id)/dp(r\,|\,c)=\frac{1}{2\lambda}\exp\left\{\frac{r^2}{2}(2\lambda-1)\right\} \qquad (2.2.7)$$

如果概率比率 LR 大于某个截断值 L，就可以认为该可见光源为射电源的真实对应体。显然(2.2.7)式仅在 $2\lambda<1$ 时作为证认的标准是有意义的。对大多数源而言，$2\lambda\approx10^{-2}$ 。

上面给出了关于单个的射电源的可见光证认的定量标准。但为了做统计研究，通常对一个样本进行证认，因此定量地估计一个可见光证认样本的可靠性和完备性是相当重要的。这里将样本的可靠性和完备性与单个源的证认概率比率联系起来。

由复合概率定律可得：

$$p(id\,|\,r)=p(id)\cdot dp(r\,|\,id)/dp(r) \qquad (2.2.8)$$

$$p(c\,|\,r)=p(c)\cdot dp(r\,|\,c)/dp(r) \qquad (2.2.9)$$

$dp(r)$ 为在以射电源的位置为中心的半径 r 到 $r+dr$ 的范围内发现可见光对应体的概率； $p(id\,|\,r)$ 与 $p(c\,|\,r)$ 为后验概率。事实上在射电源的半径 r 范围内，找到的源可能为对应体亦或混淆源，因此 $p(id\,|\,r)+p(c\,|\,r)=1$ 。$p(id)$ 为找到某个射电源的可见光对应体的先验概率。假设样本中所有的源的 Θ 部分在照相底片上存在可见光对应体，而且这些证认源是离射电源最近的源。这样 $p(id)=\Theta$ ，因而找到混淆源的先验概率为 $p(c)=1-\Theta$ 。

由概率理论的贝叶斯定理可得：

$$p(id\,|\,r)=\frac{\frac{\Theta}{1-\Theta}LR(r)}{\frac{\Theta}{1-\Theta}LR(r)+1} \tag{2.2.10}$$

$$p(c\,|\,r)=\frac{1}{\frac{\Theta}{1-\Theta}LR(r)+1} \tag{2.2.11}$$

当然，样本的可靠性和完备性依赖于概率比率的最小值 L。若概率比率小于 L，则可以排除该源不是射电源的可见光对应体。L 受两个因素制约：首先，由于不愿错过许多真的对应体，也不想得到一个十分不完备的样本，L 则不能取得太大；其次，为了使虚假的证认数目尽可能地小，L 则不能取得小，这样可靠性会随之提高。在他们的样本中，取了一个很好的折中值 L=1.8。

样本的完备性可由下式求解：

$$C=1-(\sum_{LR_i<L}p_i(id\,|\,r))/N_{id} \tag{2.2.12}$$

这里的 N_{id} 为预期的真正的对应体的总数，可以通过 $p(id\,|\,r)$ 对所有的源叠加求得。

证认样本的可靠性由下式给出：

$$R=1-(\sum_{LR_i>L}p_i(c\,|\,r))/N_c \tag{2.2.13}$$

这里的 N_c 为混在我们认为是真的对应体中的虚假源的总数，可以通过 $p(c\,|\,r)$ 对所有的证认源叠加求得。

具体概率比率的最小值L的选择可以通过做出完备性C和可靠性R随L的变化曲线，从而得到$(R+C)/2$的最大值，对应的 L 值即为要选择的值。

Sutherland等人[11]则提出了一个通常的精确的计算概率比率的方法用于源的证认。只要源的星等与位置误差分布从别的研究工作中已知，或者从控制样本的数据中估计出，这个方法即为优化的计算源证认的方法。在许多情况下，其明显优越于粗糙的最近邻天体证认的方法，尽管后一种方法似乎无偏差，但实际上对应于特殊的选择前提。他们还发展了一个自恰的公式来确认证认的可靠性。当前提条件已知的情况下，该方法可以准确地处理一对多的候选体的情况。与前人工作的不同点：在概率 $p(X\,|\,A,B)$ 有意义时，计算概率 $p(X\,|\,A)$。他们强调该方法仅作为光谱证认的参考，但不可替代光谱证认。利用可靠性估计可以优化证认的效率，减少昂贵的、难得的望远镜观测时间的浪费。Rutledge等人[12]在考虑源的星等及背景源的情况下发展了一种计算证认概率的方法，既可以计算一对一的对应体的概率，又可计算一对二的对应体的概率。以B星等和源的相似性为基础，

考虑ROSAT亮源表的源的位置不确定性，他们定量地交叉证认了ROSAT亮源表与USNO A-2 星表，给出了在X射线源的 75 角秒范围内的每一个可见光源的一对一的概率，并且给出了在三种不同的概率下ROSAT亮源表与USNO A-2 星表交叉证认出的星表，这是第一个ROSAT亮源的一对一的对应体星表，该表提供了每一个X射线源的一对一证认的概率。而且该表也包含了以USNO A-2 源为中心 10 角秒范围内对应的SIMBAD数据库中源的类型，有助于源的证认与分类。有关交叉证认的一些方法和经验可参考Egret的文章[13]。

参 考 文 献

[1] Condon J J, et al., 1996, NCSA Astronomy Digital Image Library, 1

[2] Thompson D J, et al., 1993, ApJS, 86, 629

[3] Richter G A, 1975, Astron. Nachrichten, 296, 65

[4] de Ruiter H R, Willis A G, Arp H C, 1977, A&AS, 28, 211

[5] Prestage R M, Peacock J A, 1983, MNRAS, 204, 355

[6] Wolstencroft R D, Savage A, Clowes R G, et al., 1986, MNRAS, 223, 279

[7] Willis A G, de Ruiter H R, 1977, A&AS, 29, 103

[8] Wu X B, Li Q, Zhao Y, et al., 1997, A&A, 327, L13

[9] Mattox J R, Schachter J, Molnar L, et al., 1997, ApJ, 481, 95

[10] Masci F J, Condon J J, Barlow T A, et al., PASP, 2001, 113, 10

[11] Sutherland W, Saunders W, MNRAS, 1992, 259, 413

[12] Rutledge R E, Brunner R J, Prince T A, et al., 2000, ApJS, 131, 335

[13] Egret D，Mining the Sky, Proceedings of the MPA/ESO/MPE Workshop held at Garching, Germany, 31 July-4 August, 2000. Edited by A J Banday, S Zaroubi, and M Bartelmann. Heidelberg: Springer-Verlag, 2001, p.656

§2.3 多波段数据研究方法

随着天文学正在经历的革命性的变化，“数据雪崩”这个新名词已充斥了整个天文学界。如何应付这突如其来的变化，这是摆在天文学家面前的不可回避的问题。典型的数组含有 10^8-10^9个源，每一个源具有 100 个甚至更多个测量属性，也即约 10^9个数据矢量分布在 100 维的参数空间中。从理论上而言，分析如此庞大的数组属于多变量统计分析的问题。换言之，我们需要在多维参数空间中研究天体的分布。显然，过去的那种由几个专家来分析或分类天体的做法已不再凑效。因而我们需要发展高效的、自动的分析方法，并将专家知识融入到计算机程序中，这样获得的程序利用定量的标准，避免手动方法的人为性，而且可以自动地处理大样本数据，从而可以大大地提高工作效率。正是在这种情况下，我们提出了两种自动的分类方法：支持矢量机和学习矢量量化，来研究天体的分类。通过来自三个波段（ROSAT亮源表与弱源表、USNO-A2.0、2MASS）的数据的位置交叉证认，获得了多波段数据。我们的主要目的是考察提出的方法是否适用，因此，我们选用已知样本（包括各种活动星系核、恒星和正常星系）与已获得的多波段数据交叉证认以得到训练样本和检验样本，直接运用这两种方法分类。鉴于多参数问题，我们又设想了另一种方案。为了提高分类效率，我们需要考察选出的各种参数之间是否存在相关性。若存在，则需要降维方法，如主分量分析方法或其他处理方法进行预处理。然后在降了维的空间中探测天体的分布。具体方案如图 2.2 所示，先用主分量分析方法对数据预处理，去掉那些无关的、不重要的参量，然后将处理过的数据作为支持矢量机和学习矢量量化的输入，其中一部分用以训练算法以得到分类器，另一部分用以检测获得的分类器的优劣。如果分类器的分类效率好，则可以将其用以分类新数据。通过计算结果，我们比较支持矢量机和学习矢量量化两种算法及其与主分量分析方法的混合算法的优劣。

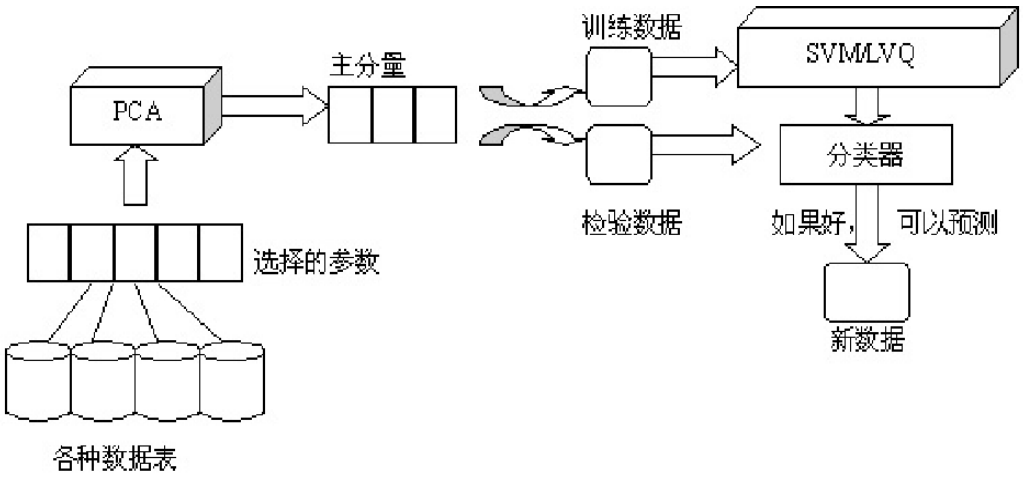


图 2.2 算法流程图

§2.4　样本的选择和参数选择

§2.4.1　样本的选择

下面以 ROSAT 亮源表与弱源表、USNO-A2.0、2MASS、Veron (2000)、SIMBAD 和 RC3 为例，介绍通过交叉证认获得样本的过程。

1990 年 6 月 1 日，德国的X射线天文台ROSAT开始了它的使命，开创了X射线天文学的新篇章。装备了大型的成像望远镜的ROSAT为天文科学提供了巨大的新的数据和洞察力。ROSAT的全天巡天亮源表[1]是在 1990/91 年ROSAT执行任务的前半年得到的。其含有18,811 个源，处于在0.1-2.4 keV的能量波段且ROSAT的PSPC的极限计数率为 0.05 cts/s。该表典型的位置误差为 30 角秒。表中列有ROSAT源的名字、赤道坐标、坐标的误差、源的计数率(CR)及其误差、背景计数率、曝光时间、硬度比HR1 和HR2 及其误差、源的扩展性（ext)、源的扩展性的可能性(extl)和源探测的可能性。类似地，ROSAT的全天巡天弱源表是亮源表的弱源扩充。表的结构与亮源表相同，含有 105,924 个弱源。其中HR1 和HR2 代表X射线的颜色，具体定义如下：

$$HR1 = \frac{B - A}{B + A} \qquad HR2 = \frac{D - C}{D + C}$$

这里A：在 0.1-0.4keV能量波段的计数率，B：在 0.5-2.0keV能量波段的计数率，C：在 0.5-0.9keV的计数率，D：在 0.9-2.0keV的计数率。

USNO-A2.0 是美国海军天文台编辑的恒星星表，含有全天 526,280,881 个恒星。其中恒星的极限星等可到 20 星等，典型的测量误差为 1 角秒。表中含有 R 星等和 B 星等。它们的有效星等范围从约 0 星等到 22 星等，并且具有较大的误差，有时高达 2 个星等，甚至更大。

2MASS星表包含近 3 亿颗恒星、50 万星系和星云在三个波段的天体测量和测光属性，以及多于 1 百万的图像数据。该表点源的测量精度准确到 0.1-0.2 角秒，展源则准确到 0.3 角秒。表中包括三个星等J(1.25μm)、H(1.65μm)和K_s(2.17μm)，它们的不确定性分别约为 0.001 个星等。

对于活动星系核，我们采用Veron（2000）的活动星系核表[2]，该表中包含了 13214 个类星体、462 个BL Lac天体和 4428 个活动星系（其中 1711 个为Seyfert 1）。恒星和一部分正常星系采自SIMBAD星表，另一部分正常星系取自第三次亮星系表RC3[3](RC3)。SIMBAD星表是由斯特拉斯堡CDS创建和维护的，其收集了大量有关 100 万河内和河外天体如恒星、星系和非恒星天体的基础数据、330 万个交叉证认数据、150 万个观测测量数据、140 万个参考文献。

为了研究已证认的 X 射线源在多维参数空间中的聚类特性就需要交叉证认星表以获得各种天体的多波段参数。为了保证尽可能在任何一个星表中至少有一

个天体对应，我们所用的交叉证认的标准和步骤如下，也可参看图 2.3，图中 XO 数据表示来自 X 射线和可见光波段的数据，XOR 数据来自 X 射线、可见光和红外波段的数据：

⑴ 首先以 RASS 源的位置为中心，将 RASS 的亮源表和弱源表与 USNO-A2.0 表在 RASS 源的 3 倍位置误差半径内交叉证认。从 RASS 的亮源表和弱源表中选取参数 CR、HR1、HR2、ext 和 extl；从 USNO-A2.0 表中选取 B 星等和 R 星等。

⑵ 然后以USNO-A2.0 源的位置为中心，将交叉证认的结果与 2MASS表在 10 角秒半径内交叉证认。从 2MASS表中选取J(1.25μm)、H(1.65μm),和K_s(2.17μm)星等。

⑶ 最后，我们以已知源的位置为中心，把活动星系核表、SIMBAD 表和 RC3 表进一步与⑵步的证认结果在 5 角秒的半径内交叉证认。从而获得已知类型样本的多波段数据。

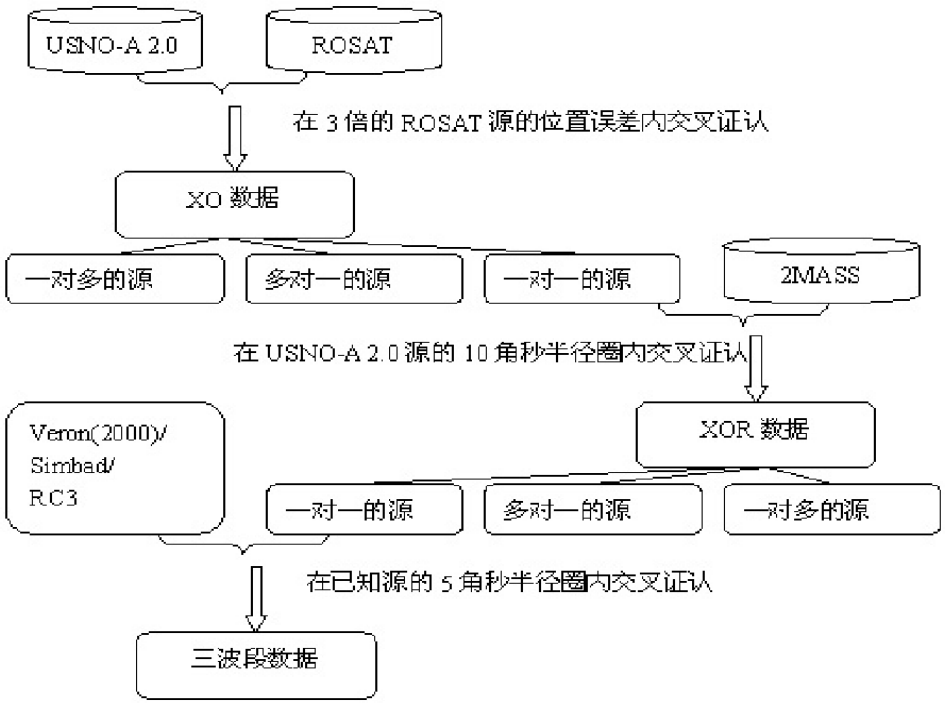


图 2.3　交叉证认的流程图

将有星表对应体的源分为三类：(i) 一对一的源；(ii) 一对多的源；(iii) 同一对应体的源。一对一的源表示在各个星表中只有一个对应体；一对多的源表示在各个星表中有多于一个源与之对应；同一对应体的源表示不同类型的天体有相同的对应体。在整个过程中，我们只考虑一对一的源。至于一对多的源，则需要通过计算交叉证认的概率方法可确认其是否为真的对应体。这样对于 X 射线和可见光波段，我们得到 2272 个类星体、336 个 BL Lac 天体、1483 个活动星系、

9967 个恒星（包括变星和白矮星）和 484 个正常星系（来自 SIMBAD 星表）；对于 X 射线、可见光和红外波段，得到 909 个类星体、135 个 BL Lac 天体、612 个活动星系、3718 个恒星（包括变星和白矮星）和 173 个正常星系（来自亮星系表 RC3）。为更清楚起见，将采纳的样本及对应的星表如表 2.1：

表 2.1　样本及其对应的星表

天体类型	两/三个波段样本数	星表
类星体	2272/909	Veron (2000)
BL Lac 天体	336/135	Veron (2000)
活动星系	1483/612	Veron (2000)
恒星	9967/3718	SIMBAD
星系	484/173	SIMBAD/RC3

§2.4.2　参数的选择

Stocke等人[4]的研究结果表明不同天体的X射线与可见光的流量比明显不同。Motch等人[5]认为为了天体的分类，最有趣的参数是各个能量波段的流量比，包括X射线硬度比、F_X/F_o流量比以及可见光颜色。在ROSAT的X射线巡天中恒星和活动星系核在数目上占主导，依靠可见光流量信息很容易将它们分开。建立在可见光参量、X射线波段的特征量如硬度比、展源特性和红外波段参量基础上，对天体的分类是合理的[4-7]。既然不同种类的天体在一些参数空间中难免重叠，所以要想将天体无偏差的证认，单取一个波段的数据是不可能的。为了将天体分类，我们取来自可见光、X射线和红外波段的数据。对不同的波段，选择的参数为B-R（可见光色指数）、B+2.5log(CR)（可见光-X射线色指数）、CR（X射线参数）、HR1（X射线参数）、HR2（X射线参数）、ext（X射线参数）、extl（X射线参数）、J-H（红外色指数）、H-K_s（红外色指数）、J+2.5log(CR)（红外-X射线色指数）。Motch等人[5]假设在 0.1-2.4 keV的能量波段，对每秒 10^{-11}尔格/平方厘米的流量与计数率的能量转换因子为每秒 1 PSPC的计数率的情况下，定义X射线与可见光的流量比：$\log(F_X/F_o)=\log(CR)+V/2.5-5.63$。因此B+2.5log(CR)和J+2.5log(CR)看作是可见光-X射线色指数和红外-X射线色指数。

我们将所得的样本的参数进行统计如表 2.2，第一列和第二列分别为参数的序号和名称，其它列为活动星系核（包括类星体、BL Lac 天体、活动星系）、恒星和正常星系的平均值及标准偏差。

表 2.2　各类天体的参数的平均值及其标准偏差

序号	参数名称	类星体	BL Lac 天体	活动星系	活动星系核	恒星	正常星系
1	B-R	0.11 ±0.51	0.78 ±0.91	0.78 ±0.89	0.41 ±0.78	-1.53 ±4.19	1.42 ±1.49
2	B+2.5log(CR)	13.87 ±1.09	15.18 ±1.57	13.02 ±2.40	13.66 ±1.83	4.18 ±5.33	7.95 ±2.40
3	CR	0.13 ±0.30	0.45 ±0.75	0.25 ±0.47	0.20 ±0.43	0.12 ±0.42	0.08 ±0.13
4	HR1	0.03 ±0.54	0.23 ±0.46	0.16 ±0.51	0.09 ±0.53	0.09 ±0.53	0.65 ±0.37
5	HR2	0.14 ±0.45	0.17 ±0.32	0.14 ±0.36	0.14 ±0.41	-0.02 ±0.54	0.22 ±0.48
6	ext	5.06 ±8.80	10.08 ±11.61	7.26 ±9.68	6.28 ±9.52	4.21 ±9.72	16.11 ±32.12
7	extl	1.15 ±4.49	5.38 ±15.23	2.20 ±5.83	1.88 ±6.62	1.05 ±6.74	7.81 ±31.15
8	J-H	0.68 ±0.27	0.75 ±0.14	0.79 ±0.15	0.73 ±0.23	0.24 ±1.77	0.76 ±0.17
9	H-K_s	0.79 ±0.31	0.70 ±0.17	0.75 ±0.23	0.76 ±0.27	0.09 ±0.11	0.37 ±0.19
10	J+2.5log(CR)	12.87 ±0.96	13.54 ±1.42	12.54 ±1.53	12.80 ±1.27	4.33 ±1.80	9.75 ±1.54

由该表我们发现不同种类的天体的各参数的平均值各不相同。很显然，正常星系的B-R值大于活动星系核和恒星的，这说明活动星系核和恒星在可见光波段

要比正常星系蓝。在活动星系核中，BL Lac天体和活动星系的B-R值大于类星体，这主要是由于类星体的颜色一般比BL Lac天体和活动星系蓝。尽管活动星系核内部各类之间有差异，但是总体上它们的各参量的平均值保持一致，明显不同于恒星和正常星系。这也意味着各类活动星系核的X射线、红外和可见光连续谱的发射机制的基本物理条件类似。而且它们具有类似的参数值支持了各种活动星系核具有相同的物理结构的假设。HR2 可以划分硬的X射线波段，对不同种类的天体的HR2 的平均值各不相同。正常星系的HR2 的平均值在某种程度上大于活动星系核和恒星的。我们也发现活动星系核和恒星具有相对较低的HR1 的平均值，也即它们具有较软的能谱分布。而正常星系则具有较硬的能谱分布，HR1 的平均值为 0.65±0.37，这是因为正常星系的相对较硬的谱来自温度高达 10^7-10^8K的热等离子体的热轫致辐射[8]。正常星系的ext和extl值大于活动星系核和恒星，这似乎由于正常星系为展源，而活动星系核和恒星为点源。活动星系核的ext和extl值也大于恒星的。正常星系的J-K_s值显然红于 0.9 星等，而且其H-K_s大于 0.2 星等，这明显红于大多数的恒星。同样活动星系也红于恒星。相比于恒星和正常星系，活动星系核的CR、B+2.5log(CR)和J+2.5log(CR)值都较大，这说明活动星系核为强的X射线发射源。

为进一步考察各类天体的区别，我们画出了各个参量的分布直方图（如图2.4），横坐标表示各个参量，纵坐标代表源的数目。图中的实线代表活动星系核，虚线代表恒星与正常星系。为简化起见，AGN代表活动星系核，S&G表示恒星与正常星系。从B-R的分布可以看出，恒星与正常星系分布范围从-12.0 到 8.0，显然宽于活动星系核的范围从-4.0 到 4.0。恒星与正常星系有两个峰值在大约-7.5 和 0.5，而活动星系核的峰值在 0.5 处。一部分恒星与正常星系比活动星系核蓝主要是由于活动星系核尤其是类星体离我们较远所致。恒星与正常星系的B+2.5log(CR)分布于-7.0 到 20.0 之间，而活动星系核则在 3.8 到 20 之间。恒星与正常星系具有两个峰值在大约在-3.6 和 5.8 处，而活动星系核的峰值为 14.0。恒星与正常星系的log(CR)分布类似于活动星系核的分布，只是高度有所不同。对于HR1 分布, 大多数的恒星与正常星系占据了HR1 较大的区域，也就是说，大多数的恒星与正常星系的谱要比活动星系核的谱硬。活动星系核在区间-0.6 到+1.0 之间呈现出较平坦的分布，而在HR1 小于-0.6 时突然下降，这可能由于星系吸收和X射线连续谱的内禀弥散造成的。在HR1 分布图上，恒星与正常星系的峰值在HR1=-0.05 和HR1=0.95 处。硬度比HR2 的直方图进一步肯定了恒星与正常星系是较硬的源。从ext和extl的分布图上可以看出，活动星系核与恒星和正常星系的分布类似，只是恒星和正常星系的高度大点儿而已。对于J-H分布和H-K_s分布，恒星与正常星系在 0.15 处有一峰值，而活动星系核的峰值在 0.75 处。相比于恒

星与正常星系，活动星系核具有较大的J-H值和H-K_s值，这表明活动星系核较红。类似于B+2.5log(CR)分布图，在J+2.5log(CR)分布图上恒星与正常星系的峰值在3.5 处，而活动星系核在 13.5 处。通常活动星系核具有较大的B+2.5log(CR)值和J+2.5log(CR)值，这与活动星系核是强的X射线发射源的事实相符。

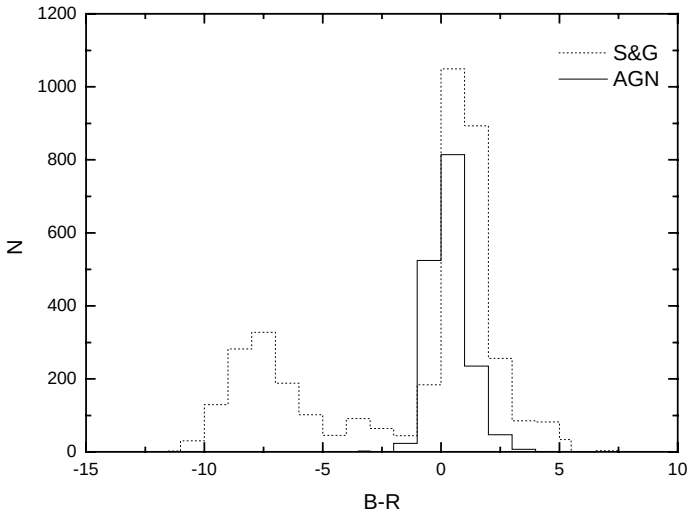


图 2.4a　B-R 直方图分布

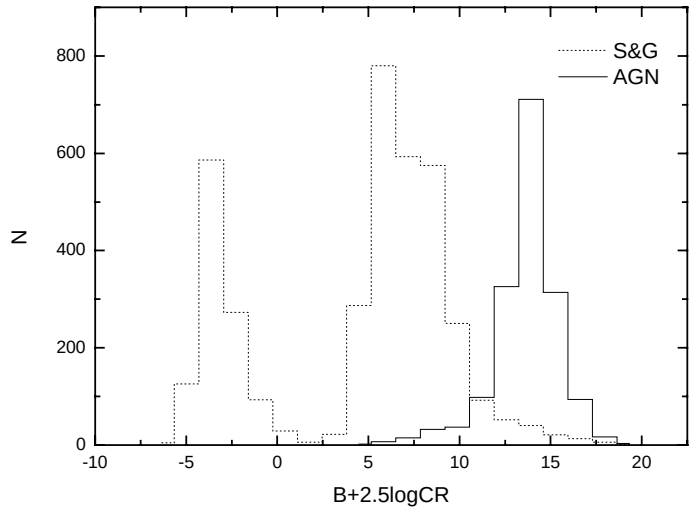


图 2.4b　B+2.5log(CR)直方图分布

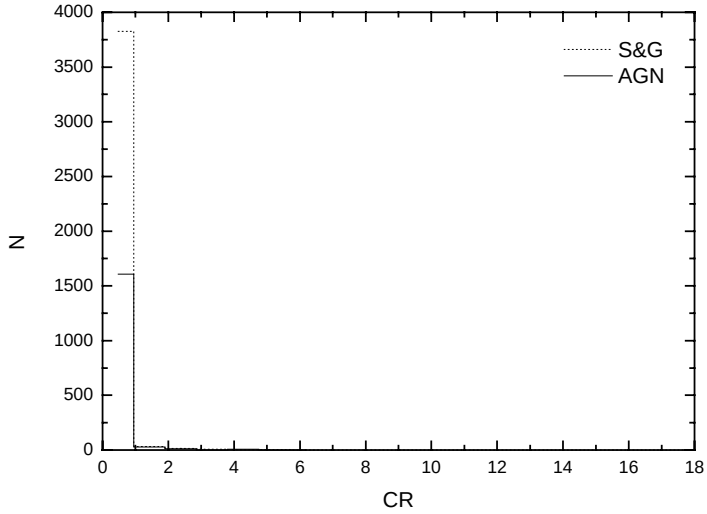


图 2.4c　CR 直方图分布

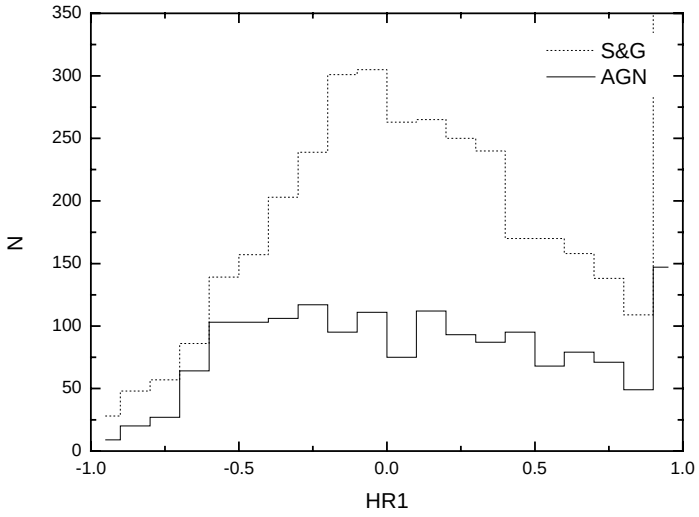


图 2.4d　HR1 直方图分布

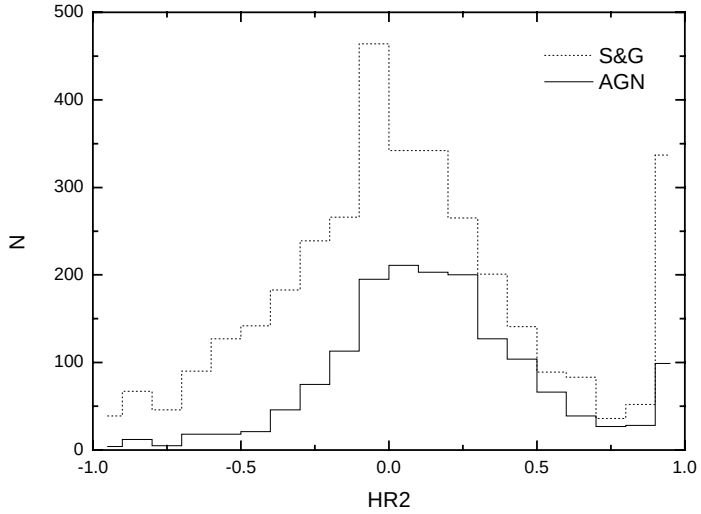


图 2.4e　HR2 直方图分布

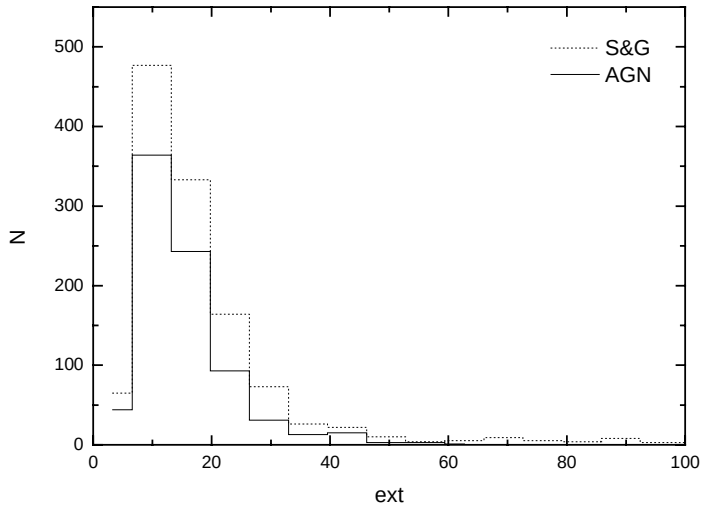


图 2.4f　ext 直方图分布

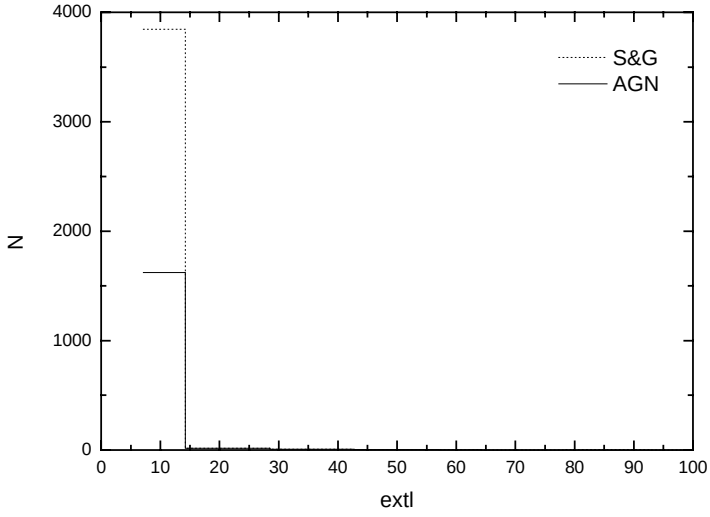


图 2.4h　extl 直方图分布

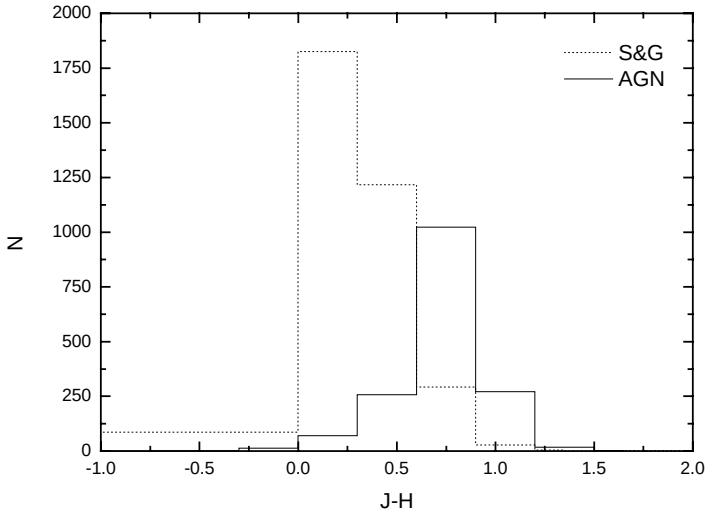


图 2.4i　J-H 直方图分布

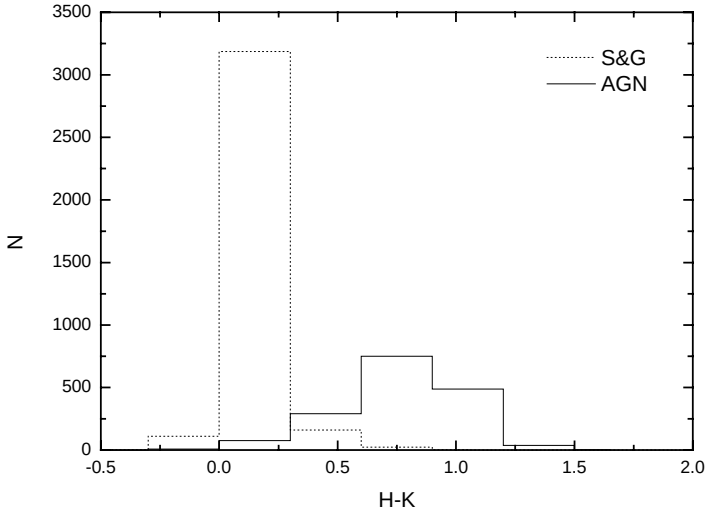


图 2.4j　H-K_s直方图分布

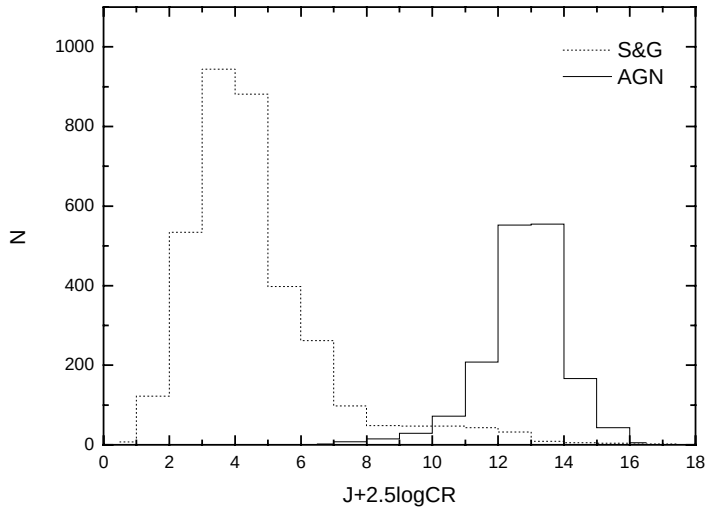


图 2.4k　J+2.5log(CR)直方图分布

通过表 2.2 和图 2.4，我们发现对不同种类的天体，来自不同波段的参数表现出不同程度的差异，因此我们考虑可以用这些参数来对天体进行分类。为了考察那些参数对分类有效，我们将在十维和低于十维的空间中探索各天体的分布情况。

参 考 文 献

[1] Voges W, Aschenbach B, Boller Th, et al., 1999, A&A 349,389

[2] Veron-Cetty M P, Veron P, 2000, ESO Scientific Report 19

[3] de Vaucouleurs G, de Vaucouleurs A, Corwin H G, et al. 1991, Third Reference Catalogue of Bright Galaxies (RC3), New York: Springer-Verlag

[4] Stocke J T, Morris S L, Gioia I M, et al., 1991, ApJS, 76, 813

[5] Motch C, Guilout P, Haberl F, et al. 1998, AASS, 132, 341

[6] Pietsch W, Bischoff K, Boller Th, et al., 1998, A&A 333, 48

[7] He X-T, Wu J-H, Yuan Q-R, et al. 2001, AJ, 121, 1863

[8] Bohringer H, 1996, In: Zimmermann H U, Trümper J E, Yorke H (eds.) MPE Report 263, Rontgenstrahlung from the Universe. Garching, P.537

第三章　支持矢量机

§3.1　支持矢量机

支持矢量机（Support Vector Machines, SVM）是一种基于统计学习理论的一般性构造学习方法，其理论是 Vapnik 于 1995 年提出，其主要思想为：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的数据。优点在于优化对偶理论使高维特征空间中的模型参数易于计算，并且运算的复杂度与问题的维数关系不大。

支持矢量机利用结构风险最小化的原理，采用最小的VC维数（Vapnik-Chervonenkis Dimension）创建分类器。若VC维数很低，误差概率会很小，这意味着有较好的推广性。用线性分割的超平面构造分类器。而对一些问题在原始空间中是线性不可分的情况，其将原始空间非线性地转换到更高维的特征空间中去。在这个特征空间中，其很容易找到一个最优的线性分割平面，即相对于训练样本，分类器具有最大的分界面[1-4]。

图 3.1　两类分类问题，“-”为一类，“+”为一类

考虑线性可分的两类分类问题如图 3.1，“-”为一类，“+”为一类。假设存在训练样本 $(x_1, y_1),...,(x_l, y_l), x \in R^n, y \in \{+1,-1\}$, l 为样本数，n 为输入维数，在线性可分的情况下就会有一个超平面 $(\omega \cdot x) + b = 0$ 使得这两类样本完全分开，该超平面满足条件：

$$(\omega \cdot x_i) + b > 0, \qquad \text{对于 } y_i = +1$$

$$(\omega \cdot x_i) + b < 0, \qquad \text{对于 } y_i = -1$$

这等价于（考虑不同的 ω 和 b），此处的“.”代表标积：

$$(\omega \cdot x_i) + b \geq 1, \qquad \text{对于 } y_i = +1 \qquad (3.1.1)$$

$$(\omega \cdot x_i) + b \leq -1, \qquad \text{对于 } y_i = -1 \qquad (3.1.2)$$

也可表示为

$$y_i[(\omega \cdot x_i)+b] \geq 1, i=1,...,l \qquad (3.1.3)$$

如果训练数据可以无误差地被分开，而且每一类数据离超平面最近的向量与超平面之间的距离最大，则称这个超平面为最优超平面。如图 3.2 所示，显然横着的平面即为最优超平面。

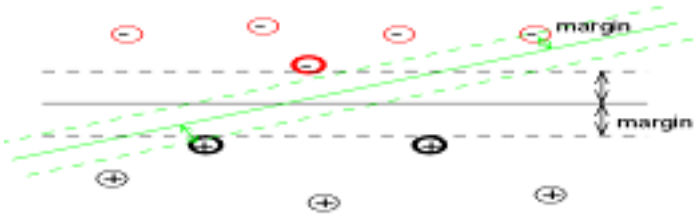


图 3.2　横的分界超平面即为最优超平面，其中实线表示分界超平面，虚线表示分类边界

为得到最优超平面，我们需要找到满足上述条件的超平面，最大化超平面与任意类训练样本的最小距离或最大化分类边界距离。处于最大的分类边界上的点为支持向量。离超平面距离最近的两个点到超平面的距离之和为：

$$\rho(\omega,b)=\min_{\{x_i|y_i=1\}}\frac{\omega\cdot x_i+b}{|\omega|}-\max_{\{x_i|y_i=-1\}}\frac{\omega\cdot x_i+b}{|\omega|} \qquad (3.1.4)$$

要想使（3.1.4）式的值最大，由（3.1.1）式可得：

$$\rho(\omega,b)=\min_{\{x_i|y_i=1\}}\frac{1}{|\omega|}-\max_{\{x_i|y_i=-1\}}\frac{-1}{|\omega|} \qquad (3.1.5)$$

$$\Leftrightarrow \rho(\omega,b)=\frac{1}{|\omega|}-\frac{-1}{|\omega|}$$

$$\Leftrightarrow \rho(\omega,b)=\frac{2}{|\omega|}$$

$$\Leftrightarrow \rho(\omega,b)=\frac{2}{\sqrt{\omega\cdot\omega}} \qquad (3.1.6)$$

因此，求解最优超平面，即为相对于矢量 ω 和标量 b，求解下式的最小值

$$\phi(\omega)=\frac{1}{2}\omega\cdot\omega=\frac{1}{2}\|\omega\|^2 \qquad (3.1.7)$$

优化函数 $\phi(\omega)$ 为二次型，约束条件是线性的，因此这是个典型的二次规划问

题，可由拉格朗日乘子法求解，引入拉格朗日乘子$\alpha_i \geq 0, i = 1,2,...,l$：

$$L(\omega,b,\alpha) = \frac{1}{2}\omega\cdot\omega - \sum_{i=1}^{l}\alpha_i\{[(x_i\cdot\omega)+b]y_i - 1\} \tag{3.1.8}$$

相对于矢量ω和标量b，L取最小值，此时的矢量ω和标量b分别表示为ω_0和b_0；相对于拉格朗日乘子α_i，L取最大值，α_i记为α_i^0。L的极值点称为鞍点。对L求导

$$\frac{\partial L(\omega,b,\alpha)}{\partial b} = 0 \tag{3.1.9}$$

$$\Leftrightarrow \sum_{i=1}^{l}\alpha_i^0 y_i = 0 \tag{3.1.10}$$

$$\frac{\partial L(\omega,b,\alpha)}{\partial \omega} = 0 \tag{3.1.11}$$

$$\Leftrightarrow \omega_0 - \sum_{i=1}^{l}\alpha_i^0 x_i y_i = 0 \tag{3.1.12}$$

从而得出最优超平面的几个特性：

① 从式（3.1.10）可以得到参数α_i^0的约束方程：

$$\sum_{i=1}^{l}\alpha_i^0 y_i = 0, \qquad \alpha_i^0 \geq 0, i = 1,2,...,l \tag{3.1.13}$$

②由（3.1.12）式可得矢量ω_0是训练样本的矢量的线性叠加：

$$\omega_0 = \sum_{i=1}^{l}\alpha_i^0 y_i x_i, \qquad \alpha_i^0 \geq 0, i = 1,2,...,l \tag{3.1.14}$$

③ 在矢量ω_0的展开式中，只有那些支持矢量的参数α_i^0值不为零：

$$\omega_0 = \sum_{\sup\ portvectors}\alpha_i^0 y_i x_i, \qquad \alpha_i^0 > 0 \tag{3.1.15}$$

由 Kühn-Tucker 定理可知：最优超平面的充分必要条件是分割超平面要满足下面的条件：

$$\alpha_i^0\{[(x_i\cdot\omega_0)+b_0]y_i - 1\} = 0, \qquad i = 1,...,l \tag{3.1.16}$$

将这些结果代入L中：

$$W(\alpha) = \frac{1}{2}\omega\cdot\omega - \sum_{i=1}^{l}\alpha_i\{[(x_i\cdot\omega)+b]y_i - 1\} \tag{3.1.17}$$

$$= \frac{1}{2}\omega\cdot\omega - \sum_{i=1}^{l}\alpha_i y_i(x_i\cdot\omega) - b\sum_{i=1}^{l}\alpha_i y_i + \sum_{i=1}^{l}\alpha_i \tag{3.1.18}$$

$$= \frac{1}{2}\omega\cdot\omega - \omega\cdot\omega - 0 + \sum_{i=1}^{l}\alpha_i \tag{3.1.19}$$

$$= \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{l} \alpha_i \alpha_j y_i y_j (x_i \cdot x_j) \qquad (3.1.20)$$

上式取最大值需在非负象限中，即

$$\alpha_i^0 \geq 0, i = 1,2,...,l \qquad (3.1.21)$$

且在下面的条件下

$$\sum_{i=1}^{l} \alpha_i y_i = 0 \qquad (3.1.22)$$

得到这个问题的解，就可以建立指示函数：

$$f(x) = sign(\sum_{spportvectors} y_i \alpha_i^0 (x_i \cdot x) - b_0) \qquad (3.1.23)$$

这里 x_i 为支持向量，α_i^0 为拉格朗日乘子，b_0 为临界值：

$$b_0 = \frac{1}{2}[(\omega_0 \cdot x^*(1)) + (\omega_0 \cdot x^*(-1))] \qquad (3.1.24)$$

其中 $x^*(1)$ 是任何属于第一类的支持矢量，$x^*(-1)$ 则为任何属于第二类的支持矢量。

这种解法仅对线性可分的数据适用，而对线性不可分的数据需略加修改，即

$$0 \leq \alpha_i^0 \leq C$$

其中 C 是一个预先假设的常数。

从上面的推导，可得出最优超平面的优点：

① 在拉格朗日乘子 α_i 不为零的情况下，最优超平面主要是由支持向量来定义的。

② 最优超平面的建立不直接依赖于所处理问题的维数。

③ 最优超平面的描述也不直接依赖于所处理问题的维数。

在线性不可分的情况下，支持向量机的主要思想是将输入矢量非线性地映射到高维特征空间中，在高维空间中寻找最优超平面。如图 3.3 所示：

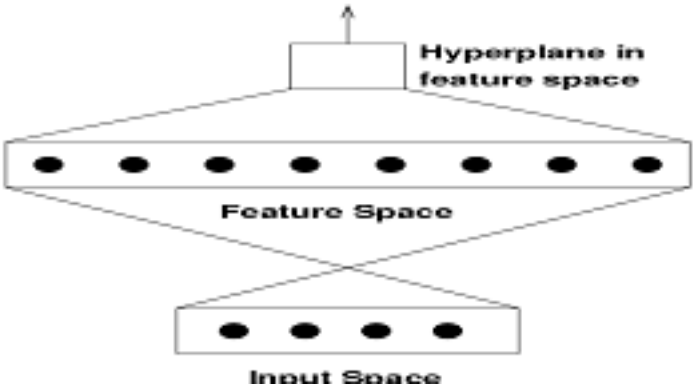


图 3.3　支持向量机将输入矢量非线性地映射到高维特征空间中

非线性映射就是将上面的标积 $x_i \cdot x$ 变为核函数 $K(x_i, x)$，这样（3.1.20）变为

$$W(\alpha) = \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{l} \alpha_i \alpha_j y_i y_j K(x_i, x_j) \qquad (3.1.24)$$

满足在下面的条件下

$$\sum_{i=1}^{l} \alpha_i y_i = 0$$

$$0 \le \alpha_i^0 \le C$$

其中 C 是一个预先假设的常数。

求解（3.1.24）式，就可以建立指示函数：

$$f(x) = sign(\sum_{spportvectors} y_i \alpha_i^0 K(x_i, x) - b_0) \qquad (3.1.25)$$

其中

$$\omega_0 \cdot x = \sum_{i=1}^{l} \alpha_i y_i K(x_i, x)$$

$$b_0 = \frac{1}{2} \sum_{i=1}^{l} \alpha_i y_i [K(x_i, x^*(1)) + K(x_i, x^*(-1))]$$

偏差 b_0 由两个支持矢量来计算。但是为了可靠性，可以用边界上的所有支持矢量计算求得。如果核函数中含有偏差项，偏差可以融入到核函数中。这样分类器会更简化：

$$f(x) = sign(\sum_{spportvectors} y_i \alpha_i^0 K(x_i, x)) \qquad (3.1.26)$$

从而简化了最优化问题。

核函数(Kernel Functions)

核函数的引入是为了建立一个到高维特征空间的映射。核函数的基本思想是使各种操作在输入空间中进行，而非在高维特征空间中进行。因此，内积并不需要在特征空间中评估。但是，计算仍旧严格地依赖于训练的样本的种类数。对于高维问题，要想获得好的数据分布通常要求足够大的训练样本。

下面的理论是建立重构带核的希尔伯特空间。特征空间的内积具有与输入空间平等的核函数，

$$K(x,x^{'})=\langle\phi(x),\phi(x^{`})\rangle\,, \qquad (3.1.27)$$

如果某种条件成立。假设 K 是一个正的对称的确定函数，且满足 Mercer 条件：

$$K(x,x^{'})=\sum_{m}^{\infty}a_m\phi_m(x)\phi_m(x^{'}), \qquad a_m\geq 0, \qquad (3.1.28)$$

$$\iint K(x,x^{'})g(x)g(x^{'})dxdx^{'}>0, \qquad g\in L_2, \qquad (3.1.29)$$

那么该核函数代表特征空间的合理内积。除非特别声明，满足Mercer条件的有效的核函数对任何x和$x^{'}$都成立。

核函数具有多种形式，这也是支持矢量机的研究热点之一。

⑴ 多项式(Polynomial)

多项式映射对于非线性模型是一种比较流行的方法，

$$K(x,x^{'})=\langle x,x^{'}\rangle^{d}\,, \qquad (3.1.30)$$

$$K(x,x^{'})=(\langle x,x^{'}\rangle+1)^{d}\,, \qquad (3.1.31)$$

为避免赫赛函数(hessian)为零，通常采用第二种。

⑵ 高斯径向基函数(Gaussian Radial Basis Function)

径向基函数备受关注，具有如下高斯形式

$$K(x,x^{'})=\exp(-\frac{\|x-x^{'}\|^2}{2\sigma^2}) \qquad (3.1.32)$$

利用径向基函数的古典技巧使用了某种确定的子族中心。典型的是一种聚类方法首先用于选择子族中心。支持矢量机的突出特征是这种选择不明显，用每一个支持矢量构造一个集中于某点的局部高斯函数。利用结构风险最小化，选择通用的基函数宽度 s 是可能的。

⑶ 指数径向基函数(Exponential Radial Basis Function)

径向基函数具有如下形式：

$$K(x,x^{'})=\exp(-\frac{\| x-x^{'} \|}{2\sigma^{2}}) \qquad (3.1.33)$$

当非连续存在时，可以得到分段的线性解。

⑷ 多层感知机(Multi-Layer Perceptron)

具有单一隐层的多层感知机有一个有效的核函数表达式：

$$K(x,x^{'})=\tanh(\rho\langle x,x^{'}\rangle+\tau) \qquad (3.1.34)$$

其中 ρ 为尺度因子，τ 为偏差因子。支持矢量对应第一层，拉格朗日对应权重。

⑸ 傅立叶级数(Fourier Series)

可以认为傅立叶级数在 2N+1 维特征空间中展开。核函数定义在$\left[-\frac{\pi}{2},\frac{\pi}{2}\right]$区间上，

$$K(x,x^{'})=\frac{\sin(N+\frac{1}{2})(x-x^{'})}{\sin(\frac{1}{2}(x-x^{'}))} \qquad (3.1.35)$$

由傅立叶转换可知，该核函数的正则能力很差，因而其不是最好的选择。

⑹ 样条函数(Splines)

由于样条函数的灵活性，它们常被采用。一个有限的κ级样条函数在τ_s处有 N 项：

$$K(x,x^{'})=\sum_{r=0}^{\kappa}x^{r}x^{'r}+\sum_{s=1}^{N}(x-\tau_s)_{+}^{\kappa}(x^{'}-\tau_s)_{+}^{\kappa} \qquad (3.1.36)$$

定义在区间[0,1)上的无限的样条函数采取如下形式：

$$K(x,x^{'})=\sum_{r=0}^{\kappa}x^{r}x^{'r}+\int_{0}^{1}(x-\tau_s)_{+}^{\kappa}(x^{'}-\tau_s)_{+}^{\kappa}d\tau \qquad (3.1.37)$$

当$\kappa=1$时，核函数变为

$$K(x,x^{'})=1+\langle x,x^{'}\rangle+\frac{1}{2}\langle x,x^{'}\rangle\min(x,x^{'})-\frac{1}{6}\min(x,x^{'})^{3} \qquad (3.1.38)$$

其解为分段的三次解。

⑺ B 样条函数(B splines)

B 样条函数是另外一种普遍的样条函数。核函数定义在[-1，1]，具体如下：

$$K(x,x^{'})=B_{2N+1}(x-x^{'}) \qquad (3.1.39)$$

⑻ 叠加的核函数(Addictive Kernels)

较为复杂的核函数可以通过核函数的叠加获得

$$K(x,x^{'})=\sum_{i}K_{i}(x,x^{'}) \qquad (3.1.40)$$

⑼ 张量积(Tensor Product)

多维核函数可由核函数的张量积得到

$$K(x,x^{'})=\prod K_{i}(x_{i},x_{i}^{'}) \qquad (3.1.41)$$

这对建立多维样条核函数尤为重要，其可直接由单变量的核函数内积得到。

很显然面对如此多的映射，究竟选取哪一个最合适？这是摆在支持矢量机研究者面前不得不面对的问题。既然一个算法中包括这么多不同的映射，这很容易比较它们的优劣。尽管核函数选择的方法的理论不断地发展，除非其在众多的问题中用独立的检测数组得以验证，否则不予采用。通常人们乐于用步步为营法(bootstrapping)和交叉确认法(cross-validation)选择核函数。

分类事例

以鸢尾属植物数据分类为例，我们看一下支持矢量机的工作原理。该数据有4个属性值，为可视化起见，我们只取最主要的两个属性，即花瓣的长度和宽度。数据分布如图3.4。考察不同的核函数及常数C对分类结果的影响，如图3.5-3.10所示。

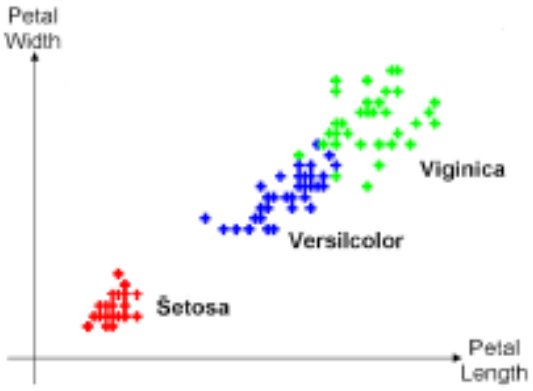


图3.4　鸢尾属植物数据分布

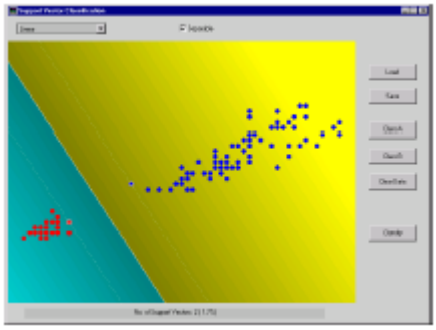


图 3.5　用线性的支持矢量机将 Setosa 数据分出来（$C=\infty$）

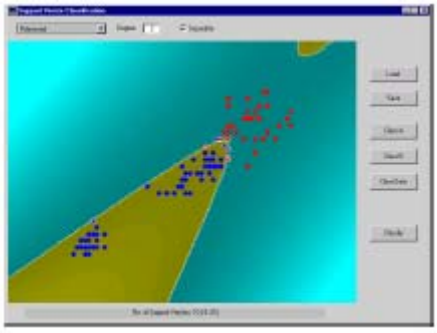


图 3.6　用 2 次多项式的支持矢量机将 Vignica 数据分出来（$C=\infty$）

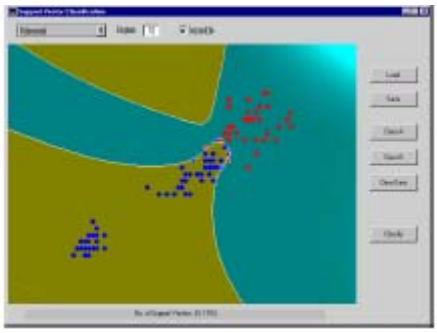


图 3.7　用 10 次多项式的支持矢量机将 Vignica 数据分出来（$C=\infty$）

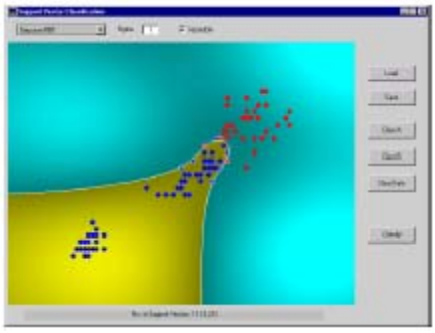


图 3.8　用径向基的支持矢量机将 Vignica 数据分出来（$\sigma=0.1, C=\infty$）

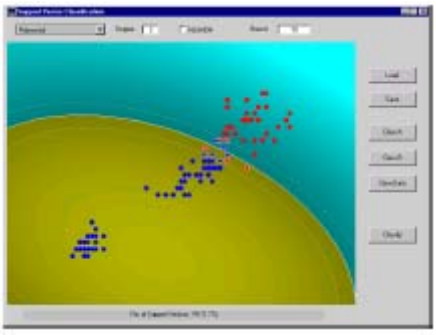


图 3.9　用 2 次多项式的支持矢量机将 Vignica 数据分出来（$C=10$）

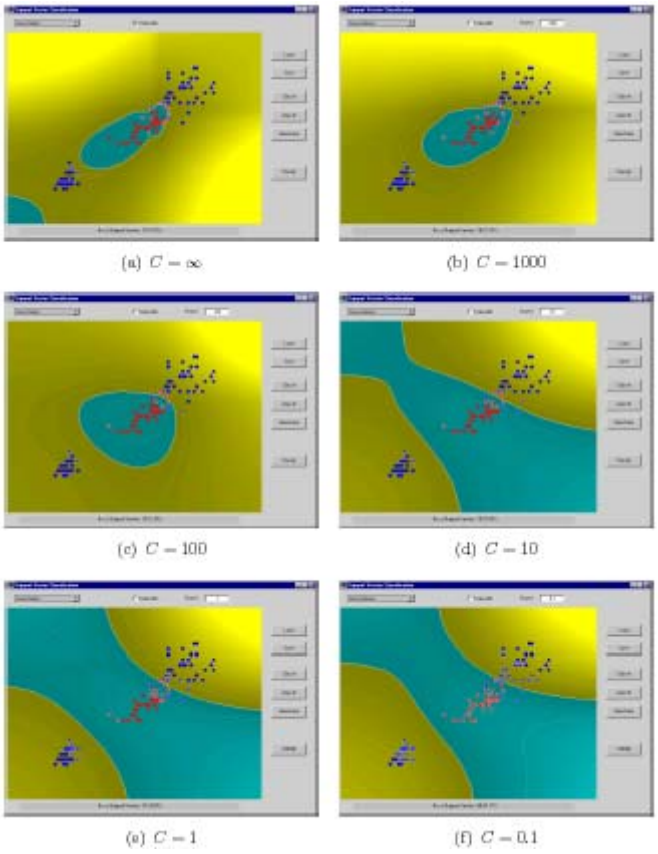


图 3.10　用线性样条的支持矢量机考察 C 对分出 Versilcolor 数据的影响

结论

对于大型的、较复杂的分类问题，应用支持矢量机来处理较合适。很突出的例子，Osuna等人[5]将其应用到人脸识别，取得了令人满意的成果。支持矢量机采用结构风险最小化的原理，减小了过渡拟合问题；利用核函数解决了维数灾难（the curse of dimensionality）问题。相比于神经网络应用的经验风险最小化，支持矢量机模型在学习效率、解决过渡拟合问题、全局优化等方面都表现出明显的优越性。结构风险最小化的原理是最小化预期风险的上限，而经验风险最小化是最小化训练样本的误差。正是基于这点不同，使得支持矢量机具有较好的推广性，

这也恰是统计学习的目的。其不仅用于两类分类问题[1, 2]和多类分类问题[6-8]，而且还用于回归[1, 2]、密度估计[9]等问题。在解决特征识别、图像压缩等问题方面也取得了一定的进展[10]。从其产生的背景和应用效果来看，该模型特别适合处理高维、复杂的目标识别问题。

参 考 文 献

[1] Gunn S R, “Support Vector Machines for Classification and Regression”, Technical Report, 1998, May, 10

[2] Stitson M O, Weston A E, Gammerman A, Vovk V, Vapnik V, “Theory of Support Vector Machines”, Technical Report CSD-TR-96-17, 1996, December, 31

[3] 史忠植，“知识发现”，清华大学出版社，2002

[4] 边肇祺，张学工等，“模式识别”，清华大学出版社，1999

[5] Osuna E R, Freund R, Girosi F, An Improved Training Algorithm for Support Vector Machines. In J Principe, L Gile, N Morgan, and E Wilson (Eds.), Neural Networks for Signal Processing VII—Proceedings of the 1997 IEEE Workshop, New York, pp. 276-285. IEEE.

[6] Weston J, Watkins C, “Multi-class Support Vector Machines”, Technical Report CSD-TR-98-04, 1998, May, 20

[7] Weston J, Watkins C, “Support Vector Machines for Multi-Class Pattern Recognition”, http://citeseer.nj.nec.com/201301.html

[8] Bredensteiner E J, “Multicategory Classification by Support Vector Machines”, http://citeseer.nj.nec.com/201301.html

[9] Weston J, Gammerman A, Stitson M, et al., “Density Estimation using Support Vector Machines”, Technical Report CSD-TR-97-23, 1998, February, 5

[10] Lothar Hermes, Dieter Frieauff, Jan Puzicha, et al. “Support Vector Machines for Land Usage Classification in Landsat TM Imagery.” In: Proc. of the IEEE International Geoscience and Remote Sensing Symposium, Hanburg, 1999: 348-350

§3.2　支持矢量机的应用

由于支持矢量机具有许多突出的吸引人的优点和很好的行为，而受到广泛的关注并被多种领域所采纳和应用。Woziak等人[1]和Humphreys等人[2]开创性地将支持矢量机应用于天文学领域。Woziak等人评估了几种自动分类方法支持矢量机（SVM）、K均值（K-means）和Autoclass用于变星分类，并对比了它们的分类效率。他们的结果表明支持矢量机在将几种确定的类别从其它样本中分离出来时表现出高效性，并且对各个类别的分类取得了较高的准确率。Humphreys等人利用不同的分类算法包括决策树、K个最近邻规则和支持矢量机对星系形态分类，并取得了较好的结果。在这里我们把支持矢量机应用到天文的多波段数据分析和处理上。

首先将来自可见光和 X 射线波段的 2272 个类星体、336 个 BL Lac 天体、1483 个活动星系、9967 个恒星和 484 个正常星系的数据作为训练样本和检验样本，我们利用支持矢量机将样本分成 5 类，分类结果如表 3.1。各类天体的正确率分别为 94.9%（类星体）、29.8%（BL Lac 天体）、19.0%（活动星系）、95.9%（恒星）、12.0%（正常星系）。对整个样本，在 14542 个源中有 2387 个误分类，占 16.4%。从该表可以清楚地看到类星体和恒星的分类结果要明显好于其它样本的分类结果。特别指出的是正常星系分类结果最差。大多数的正常星系混入恒星中，而大多数的活动星系混入到类星体中。为方便起见，我们定义活动星系核 AGNs 包括类星体、BL Lac 天体和活动星系，非活动天体 non-AGNs 包括恒星和正常星系。AGNs 具有强于 non-AGNs 的 X 射线辐射。在这两个波段，类星体和恒星的观测特性比其它天体的观测特性明显，可见整个样本是以点源为主要特征的。所以活动星核通常误分为类星体，non-AGNs 通常误分为恒星。435 个活动星系误分为恒星，这主要是因为它们具有较弱的 X 射线辐射。

类似地，将来自三个波段的数据：909 个类星体、135 个 BL Lac 天体、612 个活动星系、3718 个恒星和 173 个正常星系作为训练样本和检验样本，分类结果如表 3.2 所示。各类天体的正确率分别为 95.4%、40.0%、57.7%、99.4%、89.6%。对整个样本，5547 个源中 422 个误分占 7.6%。同样该表也表明类星体和恒星的分类结果要明显好于其它种类的分类结果。比较而言，BL Lac 天体、活动星系和正常星系的准确率较低，但是比表 3.1 有明显提高，尤其正常星系的准确率从 12.0%升至 89.6%。但是利用三个波段的数据仍不能将 BL Lac 天体从类星体和活动星系中区分出来，同时活动星系也不能明显地与类星体区分开，但是正常星系却可以很好地与恒星分开。

表 3.1　对多类分类问题，来自两个波段的样本的分类结果

分类\已知	类星体	BL Lac 天体	活动星系	恒星	正常星系
类星体	2156	152	740	355	59
BL Lac 天体	11	100	19	9	5
活动星系	41	35	282	42	21
恒星	59	47	435	9559	341
正常星系	5	2	7	2	58
正确率	94.9%	29.8%	19.0%	95.9%	12.0%

由表 3.1 尤其是表 3.2，可以看出 AGNs 明显地可与 non-AGNs 分开。我们将样本分为两类活动天体 AGNs 与非活动天体 non-AGNs：来自两个波段的 4091 个 AGNs 和 10451 个 non-AGNs；来自三个波段的 1656 个 AGNs 和 3891 个 non-AGNs。用支持矢量机对其按两类问题分类。这些样本即为训练样本又为检验样本，分类结果如表 3.3 和表 3.4。准确率对 AGNs 达 91.9%和 98.6%；对 non-AGNs 为 94.6%和 99.4%。对整个样本的准确率达 93.9%与 99.1%。

之后，我们将来自三个波段的样本分成两部分：一部分为训练样本，一部分为检测样本。我们先用 2274 个训练样本训练支持矢量机，得到分类器，然后用 2273 个检测样本测试该分类器。分类结果列入表 3.5，AGNs 与 non-AGNs 的准确率分别达 92.6%与 98.8%。在 2773 个检测样本中有 2688 个源正确分类，占 96.9%；85 个误分类达 3.1%。

表 3.2　对多类分类问题，来自三个波段的样本的分类结果

分类\已知	类星体	BL Lac 天体	活动星系	恒星	正常星系
类星体	867	49	231	7	3
BL Lac 天体	4	54	6	0	0
活动星系	37	32	353	3	12
恒星	0	0	3	3697	3
正常星系	1	0	19	11	155
正确率	95.4%	40.0%	57.7%	99.4%	89.6%

表 3.3　对两类分类问题，来自两个波段的样本的分类结果

分类\已知	AGN	non-AGN
AGN	3761	562
non-AGN	330	9889
正确率	91.9%	94.6%

表 3.4　对两类分类问题，来自三个波段的样本的分类结果

分类\已知	AGN	non-AGN
AGN	1633	25
non-AGN	23	3866
正确率	98.6%	99.4%

表 3.5　对两类分类问题，来自三个波段的样本分成两组的分类结果

分类\已知	AGN	non-AGN
AGN	767	24
non-AGN	61	1921
正确率	92.6%	98.8%

由上面的计算结果可知，我们用不同的样本检验支持向量机，分类结果表明来自三个波段的数据的分类结果明显好于来自两个波段的数据的分类结果，不论对两类分类问题还是多类分类问题。可见利用这 10 个参数B-R、B+2.5log(CR)、CR、HR1、HR2、ext、extl、J-H、H-K_s和J+2.5log(CR)将AGN从non-AGN中分离出来完全是可能的。由表 3.5 可知，AGN和non-AGN的分类准确率高达 92.0%以上，即这 10 个特征量作为将活动星系核从非活动星系核中分辨出来是有效的。10 个特征量的分类准确率明显高于 7 个特征量，即提取的特征量越多，分类效果越好。如果增加其它波段的数据例如射电波段，不仅AGN可以与non-AGN较好地分开，而且各类AGN之间也可能区分开。对于BL Lac天体和活动星系的准确率较低，可能由于样本数太少所致。随着数据在数量和质量上的提高，我们相信分类结果会更好。支持向量机的优点就是其应用了结构风险最小化的原理。利用该原理创建分类器，若结构风险最小，那么预期的误差概率则最低。这意味着

支持向量机具有较好的推广性。对一些在原始空间中不能线性分割的问题，支持向量机可以利用线性超平面创建分类器，也可以将输入矢量非线性地映射到高维空间中创建分类器。在这个高维空间中很容易找到线性最优分类面。通过支持向量机的输出可以看出分类的可靠性，这也正是支持向量机成为备受欢迎的方法的原因之一。

参 考 文 献

[1] Wozniak P R, Akerlof C, Amrose, S., et al. 2001, AAS 199, 130, 04

[2] Humphreys R M, Karypis G, Hasan M, et al. 2001, AAS 199, 10, 15

第四章　神经网络

§4.1　神经网络

神经网络（Neural Network, NN），又称为人工神经网络(Artificial Neural Network，ANN), 是指为了模拟动物神经细胞群学习特性的结构和功能而构成的一种信息处理系统或计算机系统。由于其拥有很强的适用于复杂环境和多目标控制要求的能力，并具有以任意精度逼近任意非线性连续函数的特性（自组织、自学习、自适应）而适用于复杂系统的控制的应用领域。

神经网络是由数据驱动的，这意味着它们必须输入大量关于系统过去特性的数据由其分析，称为“训练”。在“训练”期间，神经网络系统研究它收到的原始的随机性数据，重建它们的数学关系，将之转换为连续性的有数学规律性的形式，结果得到一个适合这些数据的模型（称为建模）。即神经网络系统依据被控系统的输入输出数据对，通过学习得到一个描述系统输入输出关系的非线性映射，然后神经网络系统还会自动对它本身的模型进行调整，所以神经网络系统还有“预测”功能。

常规统计学意义上的统计学模型可以用一些二次或更高阶的方程式来表述，这些方程式可以从理论上给出输入变量间的关系并得出结果。而神经网络模型无法用常规的方程式来表述，它的信息大多依存于它的结构的各点之中，依赖模型本身的多层（一般为三层）结构和自学习特性，通过“训练”和“学习”建立及调整模型，计算的重担被推给模型本身，在“训练”阶段，模型已建立了适合这些数据的内部结构，而不需要额外的程序。

神经网络目前在模式识别、机器视觉和听觉、智能计算、机器人控制、信号处理、联想记忆、数据挖掘、医学诊断、金融决策、过程控制和组合优化等领域得到广泛应用。其中，数据挖掘是神经网络目前最新的和最重要的应用领域。

神经元模型的提出

“人工神经网络”是在对人脑组织结构和运行机智的认识理解基础之上模拟其结构和智能行为的一种工程系统。早在本世纪 40 年代初期，心理学家 McCulloch、数学家 Pitts 就提出了人工神经网络的第一个数学模型，从此开创了神经科学理论的研究时代。其后，F. Rosenblatt、Widrow 和 Hopf、J. J. Hopfield 等学者又先后提出了感知模型，使得人工神经网络技术得以蓬勃发展。

神经系统的基本构造是神经元(神经细胞)，它是处理人体内各部分之间相互信息传递的基本单元。据神经生物学家的研究结果表明，人的一个大脑一般有10^{10}～10^{11}个神经元。神经元结构图如图 4.1 所示，每个神经元都是由一个细胞体、

一个连接其它神经元的轴突和一些向外伸出的其它较短分支——树突组成。轴突的功能是将本神经元的输出信号(兴奋)传递给别的神经元。其末端的许多神经末梢使得兴奋可以同时传送给多个神经元。树突的功能是接受来自其它神经元的兴奋。神经元细胞体将接受到的所有信号进行简单地处理(如：加权求和，即对所有的输入信号都加以考虑且对每个信号的重视程度体现在权值上的不同)后由轴突输出。神经元的树突与另外的神经元的神经末梢相连的部分称为突触。

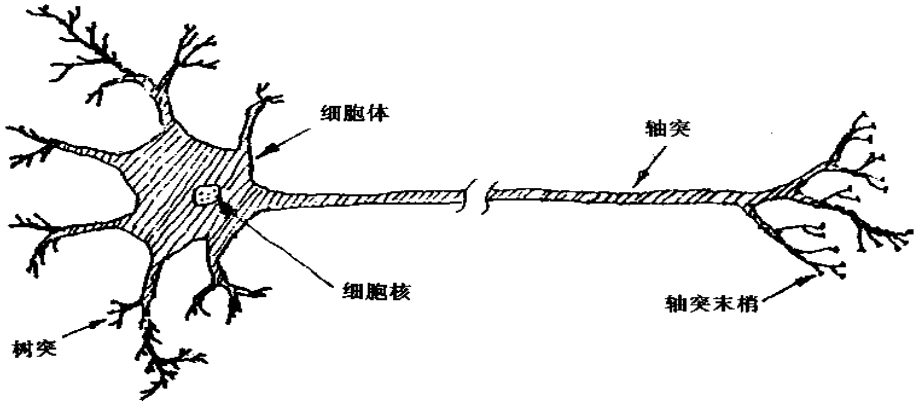


图 4.1　神经元结构图

从神经元的特性和功能可以知道，神经元是一个多输入单输出的信息处理单元，而且，它对信息的处理是非线性的。根据神经元的特性和功能，可以把神经元抽象为一个简单的数学模型。工程上用的人工神经元模型如图 4.2 所示。在图 4.2 中，X_1，X_2，……，X_n是神经元的输入，即是来自前级n个神经元的轴突的信息；θ_i是i神经元的阈值；W_{i1}，W_{i2}……，W_{in}分别是i神经元对X_1，X_2，……，X_n的权系数，也即突触的传递效率；Y_i是i神经元的输出；f[·]是激发函数，它决定i神经元受到输入X_1，X_2，……，X_n的共同刺激达到阈值时以何种方式输出。

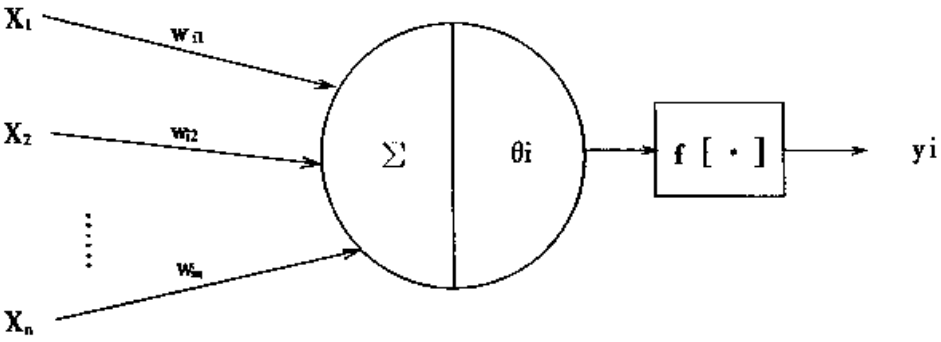


图 4.2　神经元模型

从图 4.2 的神经元模型，可以得到神经元的数学模型表达式：

$$f(u_i)=\begin{cases}1 & u_i>0\\ 0 & u_i\le 0\end{cases} \qquad (4.1.1)$$

对于激发函数 f[·]有多种形式，其中最常见的有阈值型、线性型和 Sigmoid

型三种形式，这三种形式如图 4.3 所示。

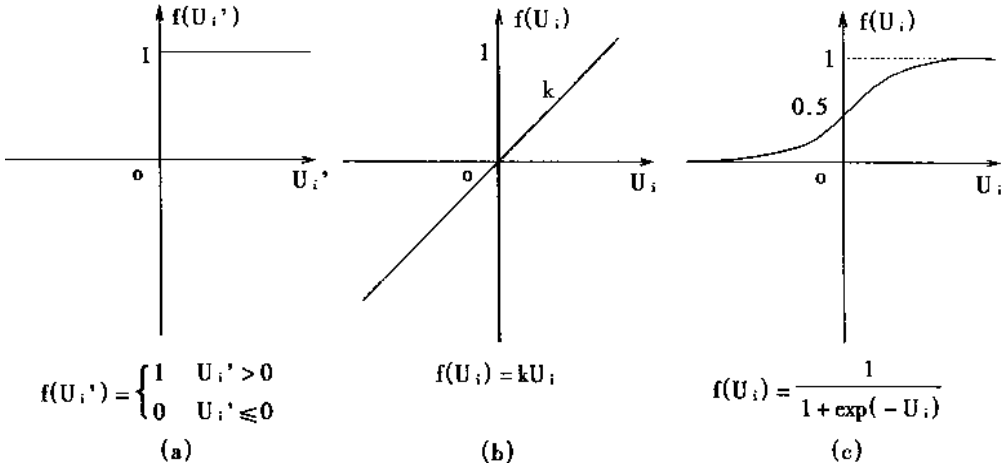


图 4.3　典型的激发函数

为了表达方便；令：
$$u_i = \sum_{j=1}^{n} W_{ij} X_j - \theta_i \qquad (4.1.2)$$

则式(4.1.1)可写成下式：
$$y_i = f(u_i) \qquad (4.1.3)$$

显然，对于阈值型激发函数有：
$$f(u_i) = \begin{cases} 1 & u_i > 0 \\ 0 & u_i \le 0 \end{cases} \qquad (4.1.4)$$

对于线性型激发函数，有：
$$f(u_i) = ku_i \qquad (4.1.5)$$

对于 Sigmoid 型激发函数，有：
$$f(u_i) = \frac{1}{1+e^{-u_i}} \qquad (4.1.6)$$

对于阈值型激发函数，它的输出是电位脉冲，故而这种激发函数的神经元称离散输出模型；对于线性激发函数，它的输出是与输入的激发总量成正比的，故这种神经元称线性连续型模型；对于用 Sigmoid 型激发函数，它的输出是非线性的，故这种神经元称非线性连续型模型。

上面所叙述的是最广泛应用而且人们最熟悉的神经元数学模型，也是历史最长的神经元模型。近若干年来，随着神经网络理论的发展，出现了不少新颖的神经元数学模型，这些模型包括逻辑神经元模型、模糊神经元模型等，并且渐渐地受到人们的关注和重视。

神经网络结构及功能

神经元和神经网络的关系是元素与整体的关系。神经元的结构很简单，工作机理也不深奥；但是用神经元组成的神经网络就非常复杂，其功能也十分奥妙。

人们平常十分清楚砖头是很简单的，但是用简单的砖头，人们就可以筑造出

各种美伦美奂的建筑物，无论是优雅的别墅、亦或是高耸雄伟的大厦、或者是粗旷的金字塔、亦或是庄严肃穆的教堂，无一不是由简单的砖头堆砌而成。简单的神经元也是如此，通过不同方式的连接和信息传递，也就能产生丰富多彩的神经网络结构，创造出令人赞叹的优异功能。

神经网络就是由许多神经元互连在一起所组成的神经结构。把神经元之间相互作用的关系进行数学模型化就可以得到神经网络模型。

⑴ 神经网络的基本属性

神经网络有些基本属性，它们反映了神经网络的特质。

① 非线性

人脑的思维是非线性的，故人工神经网络模拟人的思维也应是非线性的。

② 非局域性

非局域性是人的神经系统的一个特性，人的整体行为是非局域性的最明显体现。神经网络以大量的神经元连接模拟人脑的非局域性，它的分布存储是其非局域性的一种表现。

③ 非定常性

神经网络是模拟人脑思维运动的动力学系统，它应按不同时刻的外界刺激对自己的功能进行修改，故而它是一个时变的系统。

④ 非凸性

神经网络的非凸性即是指它有多个极值，也即系统具有不只一个较稳定的平衡状态。这种属性会使系统的演化多样化。神经网络的全局优化算法就反映了这一点，例如模拟退火法。

⑵ 神经网络的主要特征

① 学习性

神经网络的建模是通过训练样本学习得到的。

② 推广性

可以将训练样本得到的规则用于新的样本。

③ 错误容忍性

传统的计算机要求处理的数据必须精确，而神经网络即使面对有噪声的、不完整的数据，也能应付自如。

④ 快速性

因为许多相互联系的处理单元以平行方式工作，而且模拟人脑的神经网络拥有成千上亿个神经元，并不会随时间的变化而退化。

⑤ 一旦破坏，适度恶化

一旦遭到破坏，平常的软硬件将会停止工作，而神经网络则只是适度恶化。

⑥ 建立廉价，但计算上集中于训练

建立神经网络模型相对而言较容易，只是在计算上集中于训练。

⑦ 适宜于处理复杂棘手的问题

对于不能通过算法、方程、规则来表示；输入与输出存在对应关系，但具体的映射函数未知；具有足够多的训练样本的情况，都可以用神经网络来实现。

⑧ 以训练得到非线性映射

处理现实的数据，通过训练可以产生非线性映射，例如预报天气。

⑶ 神经网络的学习方法

神经网络性质的主要取决于以下两个因素：① 网络的拓扑结构；② 网络的权值、工作规则。二者合起来就可以构成网络的主要特征。

随着网络结构和功能的不同，网络权值的学习算法也不同。神经网络的连接权值确定方式一般有两种：① 通过设计计算确定，即死记式学习；② 网络按一定的规则通过学习（训练）得到的。通常采取后者确定其权值。

对于神经网络的学习方法，从学习过程的组织与管理分：有监督学习和无监督学习；从学习过程的推理和决策分：确定性学习、随机学习和模糊学习。

⑷ 神经网络模型

神经网络作为模拟复杂系统非线性关系的一种模型，按其内部神经元连接的拓扑结构、学习规则以及传递函数的类型等标准可以分为若干种类。神经网络在目前已有几十种不同的模型。通常可按如下原则进行分类：

按照网络的结构区分：前向网络和反馈网络。

按照学习方式区分：有教师学习和无教师学习网络。

按照网络性能区分：连续型和离散性网络、随机型和确定型网络。

按照突触性质区分：一阶线性关联网络和高阶非线性关联网络。

按对生物神经系统的层次模拟区分：神经元层次模型、组合式模型、网络层次模型、神经系统层次模型和智能型模型。

通常，人们较多地考虑神经网络的互连结构。一般而言，神经网络有分层网络、层内连接的分层网络、反馈连接的分层网络、互连网络等 4 种互连结构。在人们提出的几十种神经网络模型中，应用较多的是前向型神经网络、径向基函数神经网络、Hopfield 神经网络、BP 网络、Kohonen 网络和 ART(自适应共振理论)网络。

⑸ 几种典型神经网络简介

① 多层感知网络

多层感知网络又称误差逆传播神经网络，在 1986 年以 Rumelhart 和 McCelland 为首的科学家出版的《Parallel Distributed Processing》一书中，完整地提出了误差逆传播学习算法，并被广泛接受。多层感知网络是一种具有三层或三层以上的阶层型神经网络。典型的多层感知网络是三层前馈的阶层网络，即输入层、隐含层(也称中间层)、输出层。相邻层之间的各神经元实现权连接，即下一层的每一个神经元与上一层的每个神经元都实现权连接，而且每层各神经元之间无连接。

学习规则及过程：它以一种有教师示教的方式进行学习。首先由教师对每一种输入模式设定一个期望输出值。然后对网络输入实际的学习记忆模式，并由输入层经中间层向输出层传播(称为“模式顺传播”)。实际输出与期望输出的差即是误差。按照误差平方最小这一规则，由输出层往中间层逐层修正连接权值，此过程称为“误差逆传播”。所以误差逆传播神经网络也简称 BP(Back Propagation)网。随着“模式顺传播”和“误差逆传播”过程的交替反复进行。网络的实际输出逐渐向各自所对应的期望输出逼近，网络对输入模式的响应的正确率也不断上升。通过此学习过程，确定了各层间的连接权值之后就可以工作了。

由于 BP 网及误差逆传播算法具有中间隐含层并有相应的学习规则可循，使得它具有对非线性模式的识别能力。特别是其数学意义明确、步骤分明的学习算法，更使其具有广泛的应用前景。目前，在手写字体的识别、语音识别、文本与语言转换、图像识别以及生物医学信号处理方面已有实际的应用。

但 BP 网并不十分完善，它存在以下一些主要缺陷：学习收敛速度太慢、网络的学习记忆具有不稳定性，即当给一个训练好的网提供新的学习记忆模式时，将使已有的连接权值被打乱，导致已记忆的学习模式的信息的消失。

② 竞争型 Kohonen 神经网络

竞争型 Kohonen 神经网络是基于人的视网膜及大脑皮层对刺激的反应而引出的。神经生物学的研究结果表明：生物视网膜中，有许多特定的细胞，对特定的图形(输入模式)比较敏感，并使得大脑皮层中的特定细胞产生大的兴奋，而其相邻的神经细胞的兴奋程度被抑制。对于某一个输入模式，通过竞争在输出层中只激活一个相应的输出神经元。许多输入模式，在输出层中将激活许多个神经元，从而形成一个反映输入数据的“特征图形”。

竞争型神经网络是一种以无教师方式进行网络训练的网络。它通过自身训练，自动对输入模式进行分类。竞争型神经网络及其学习规则与其它类型的神经

网络和学习规则相比，有其自己的鲜明特点。在网络结构上，它既不象阶层型神经网络那样各层神经元之间只有单向连接，也不象权连接型网络那样在网络结构上没有明显的层次界限。它一般是由输入层(模拟视网膜神经元)和竞争层(模拟大脑皮层神经元，也叫输出层)构成的两层网络。两层之间的各神经元实现双向权连接，而且网络中没有隐含层。有时竞争层各神经元之间还存在横向连接。竞争型神经网络的基本思想是网络竞争层各神经元竞争对输入模式的响应机会，最后仅有一个神经元成为竞争的获胜者，并且只将与获胜神经元有关的各连接权值进行修正，使之朝着更有利于它竞争的方向调整。神经网络工作时，对于某一输入模式，网络中与该模式最相近的学习输入模式相对应的竞争层神经元将有最大的输出值，即以竞争层获胜神经元来表示分类结果。这是通过竞争得以实现的，实际上也就是网络回忆联想的过程。

除了竞争的方法外，还有通过抑制手段获取胜利的方法，即网络竞争层各神经元抑制所有其它神经元对输入模式的响应机会，从而使自己“脱颖而出”，成为获胜神经元。除此之外还有一种称为侧抑制的方法，即每个神经元只抑制与自己邻近的神经元，而对远离自己的神经元不抑制。这种方法常常用于图像边缘处理，解决图像边缘的缺陷问题。

竞争型神经网络的缺点和不足：因为它仅以输出层中的单个神经元代表某一类模式。所以一旦输出层中的某个输出神经元损坏，则导致该神经元所代表的该模式信息全部丢失。

③ Hopfield 神经网络

1986 年美国物理学家 Hopfield 陆续发表几篇论文，提出了 Hopfield 神经网络。他利用非线性动力学系统理论中的能量函数方法研究反馈人工神经网络的稳定性，并利用此方法建立求解优化计算问题的系统方程式。基本的 Hopfield 神经网络是一个由非线性元件构成的权连接型单层反馈系统。

网络中的每一个神经元都将自己的输出通过连接权值传送给所有其它神经元，同时又都接收所有其它神经元传递过来的信息。即：网络中的神经元 t 时刻的输出状态实际上间接地与自己的 t-1 时刻的输出状态有关。所以 Hopfield 神经网络是一个反馈型的网络。其状态变化可以用差分方程来表征。反馈型网络的一个重要特点就是它具有稳定状态。当网络达到稳定状态的时候，也就是它的能量函数达到最小的时候。这里的能量函数不是物理意义上的能量函数，而是在表达形式上与物理意义上的能量概念一致，表征网络状态的变化趋势，并可以依据 Hopfield 工作运行规则不断进行状态变化，最终能够达到某个极小值的目标函数。网络收敛就是指能量函数达到极小值。如果把一个最优化问题的目标函数转换成网络的能量函数，把问题的变量对应于网络的状态，那么 Hopfield 神经网络

就能够用于解决优化组合问题。

Hopfield工作时其各个神经元的连接权值是固定的，更新的只是神经元的输出状态。Hopfield神经网络的运行规则为：首先从网络中随机选取一个神经元u_i，按照公式(3.2.1)进行加权求和，再按公式(3.2.2)计算u_i的第t+1 时刻的输出值。除u_i以外的所有神经元的输出值保持不变，返回至第一步，直至网络进入稳定状态。

对于同样结构的网络，当网络参数(指连接权值和阈值)有所变化时，网络能量函数的极小点(称为网络的稳定平衡点)的个数和极小值的大小也将变化。因此，可以把所需记忆的模式设计成某个确定网络状态的一个稳定平衡点。若网络有 M 个平衡点，则可以记忆 M 个记忆模式。

当网络从与记忆模式较靠近的某个初始状态(相当于发生了某些变形或含有某些噪声的记忆模式，也即只提供了某个模式的部分信息)出发后，网络按 Hopfield 工作运行规则进行状态更新，最后网络的状态将稳定在能量函数的极小点。这样就完成了由部分信息的联想过程。

Hopfield 神经网络的能量函数是朝着梯度减小的方向变化，但它仍然存在一个问题，那就是一旦能量函数陷入局部极小值，它将不能自动跳出局部极小点，到达全局最小点，因而无法求得网络最优解。这可以通过模拟退火算法或遗传算法得以解决，在此不再一一介绍。

产生于不同起源和针对不同目的的神经网络模型有很多种，这里只主要简要地介绍了一下具有代表性的神经网络模型：多层感知器、Kohonen 神经网络、Hopfield 神经网络。前两者也是模式识别中最典型的两种模型，后者更多地侧重于组合优化问题。要想更深入地了解神经网络及其应用，则需要参考有关文献和神经网络方面的专著。

⑹ 数据的收集和预处理

另外，应用神经网络时，需认真收集数据并对数据预处理。神经网络是通过训练样本学习获得模式和规则的，故数据的质量和数量严重影响神经网络的优劣。因此数据的收集和预处理是神经网络的重中之重。数据预处理时应注意几点：

① 理想上，训练样本矢量数、证明样本矢量数以及验证样本矢量数应相等；

② 输入矢量归一化，以利于提取有效特征矢量；

③ 训练样本要完备；

④ 对分类问题，要保证各类样本的平衡性，而且对于分类和预测问题，要用到目标数据且采用监督的学习方法；

⑤ 画直方图，有助于发现离群数据（outliers）；

⑥ 对缺值问题，简便方法是用样本的平均值代替，分类情况则用类平均值

代替。

结束语

神经网络的研究内容相当广泛，反映了多学科交叉技术领域的特点。迄今为止，在人工神经网络研究领域中，有代表性的网络模型已达数十种，而学习算法的类型更难以统计其数量。神经网络研究热潮的兴起是本世纪末人类科学技术发展全面飞跃的一个组成部分。由于神经网络非常适用于处理数据的非线性复杂关系，并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，因而被广泛地应用于数据挖掘和知识发现中，并以不同的网络模型分别实现了聚类、分类、关联、回归、模式识别等多种算法。神经网络在众多领域广泛应用的同时，也遇到一些难以解决的问题。如对付数据量巨大、非线性程度很高的数学集时，神经网络存在学习速度慢、难以收敛等问题；而对于采用自组织增量式学习方式的网络，会使其结构急剧膨胀，甚至崩溃；此外，神经网络的另一个突出弱点就是当使用带有噪声的数据训练网络时，往往会因训练过程控制的不当而使网络产生过学习和欠学习现象，从而影响网络预测结果。尽管这些问题存在，而且在很大程度上制约了神经网络理论和应用的发展，不过现在人们已经充分认识到这些问题，并开始深入地研究。神经网络与多种科学领域的发展密切相关，纵观当代新兴科学技术的发展历史，人类在征服宇宙空间、基本粒子、生命起源等科学领域的进程之中历经了崎岖不平之路。我们也会看到，探索人脑功能和神经网络的研究将伴随着重重困难的克服而日新月异。

§4.1.1 自组织映射

自组织映射（Self-organization Map, SOM）是一种无监督的神经网络[4, 5]，自动地从数据中提取特征，试图发现数据中的潜在结构，这属于聚类分析。具有相似的输入矢量的数据聚集在一起，为一类。自组织映射可以较好地完成聚类的任务，其中每一个神经元节点对应一个聚类中心。与普通聚类算法不同的是，所得的聚类之间仍保持一定的关系，就是在自组织映射网络节点平面上相邻或相隔较近的节点对应的类别，它们之间的相似性要比相隔较远的类别之间大。因此可以根据各个类别在节点平面上的相对位置进行类别的合并和类别之间关系的分析。

1、自组织映射的特点和功能：

① 无监督性；

② 自动地提取特征、主分量分析、聚类、编码、特征映射；

③ 有助于可视化，将高维数据投影到二维平面上，保持拓扑映射。

2、自组织映射的算法

① 权值初始化，固定最大的拓扑半径 R 和学习率 α；

② 计算神经元的输入矢量与权重矢量之间的距离；

$$D(j)=\Sigma(\omega_{ij}-x_i)^2$$

③ 找最小值 D(J)；

④ 在确定的神经元 J 的附近，对所有神经元 j 及 i 存在

$$\omega_{ij}(new)=\omega_{ij}(old)+\alpha(x_i-\omega_{ij}(old))$$

⑤ 减小学习率 α，并在一定的时间后减小拓扑半径 R；

⑥ 当学习率 α 足够小和权重矢量不再改变时，停止训练。

§4.1.2 学习矢量量化

学习矢量量化(Learning Vector Quantization，LVQ)是相当新的一类神经网络模型[6, 7]，自 1988 年Kohonen提出此模式以来，由于这种网络的学习速度比反向传输网络快，因此颇受重视。但是它也有一个缺点：所需的隐藏层处理单元与输出的处理单元成正比，而且需整数倍于它。若与概率神经网络相比，则其学习速度较慢，但回想速度较快，且所需的记忆体较小，因此学习矢量量化(LVQ)可以说是一种介于反向传输网络与概率神经网络之间的模式。与自组织映射的关系：区别于SOM，具有监督性；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。

1、学习矢量量化网络构架(如图 4.4 所示)

① 输入层

用以表现网络的输入变量，即训练范例的输入向量（或称特征相量），其处理单元的数目以问题而定。使用线性转换函数，即 F(X)=X。

② 隐藏层

用以表现输入各类样本点群的中心坐标。每一个处理单元表示一个中心坐标，其与输入层处理单元相连的连接储存着该中心坐标。与概率神经网络相似之处，在概率神经网络中隐藏层用以表现训练范例，每个处理单元表示一个训练范例，其与输入层处理单元相连接的连接储存着该训练范例的特征向量。与概率神经网络不同之处在于其隐藏层处理单元代表一群输入向量相似，且具有相同目标输出向量（即分类）的训练范例之代表性范例，而非别的训练范例。另外每一个隐藏层处理单元将属于一个特定输出层处理单元。

③ 输出层

用以表现分类, 每个处理单元表示一个分类。通常隐藏层处理单元的数目取输出层处理单元的整数倍，例如 5 倍, 实际上这个数目反映分类其训练范例(样本代表点)散布群数的大概估计值。输出层处理单元与隐藏层处理单元相连的连接储存着隐藏层处理单元的分类信息。总之, 学习矢量量化的输入层与隐藏层间的加权值为变数, 需通过学习来决定, 而隐藏层与输出层间加权值为固定值, 不需要通过学习来决定。

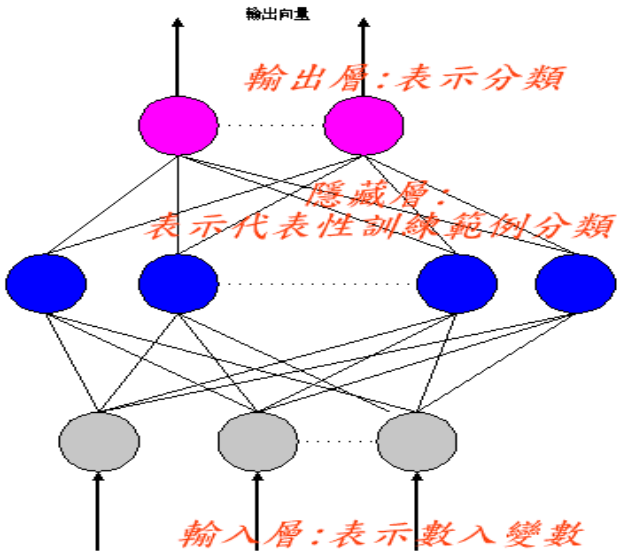


图 4.4　学习矢量量化网络构架

2、学习矢量量化的算法

① 权值和固定学习率 α 初始化；

② 计算神经元的输入矢量与权重矢量之间的欧几里得距离；

$$D(j)=(\Sigma(\omega_{ij}-x_i)^2)^{0.5}$$

③ 找最小值 D(J)；

④ 更新神经元 J 的权重如下；

如果神经元 J 的类别与输入正确类别相同：

$$\omega_{ij}(new)=\omega_{ij}(old)+\alpha(x_i-\omega_{ij}(old))$$

如果神经元 J 的类别与输入正确类别不同：

$$\omega_{ij}(new)=\omega_{ij}(old)-\alpha(x_i-\omega_{ij}(old))$$

⑤ 减小学习率 α；

⑥ 当学习率 α 足够小和权重矢量不再改变时，停止训练。

参 考 文 献

[1] 史忠植，2002，知识发现，北京：清华大学出版社

[2] 边肇祺，张学工等，1999，模式识别，北京：清华大学出版社

[3] 阎平凡，黄端旭，1993，人工神经网络——模型、分析与应用，合肥：安徽教育出版社

[4] Kohonen T, The self-organization map. Proceedings of the IEEE, 1990, 78(9): 1464-1480

[5] Kohonen T, Self-Organization Maps, Berlin: Springer-Verlag, 1995

[6] Kohonen T, New Developments of Learning Vector Quantization and the Self-Organization Map. In Symposium on Neural Networks; Alliances and Perspectives in Senri 1992(SYNAPSE'92), Osaka, Japan

[7] Kohonen T, Kangas J, Laaksonen J, Torkkpla K, LVQ_PAK: The Learning Vector Quantization Program Package, Version 3.1, April 7, 1995

§4.2　学习矢量量化的应用

我们本文采纳的学习矢量量化的算法是以芬兰的赫尔辛基科技大学的计算机与信息实验室发展的LVQ_PAK程序为基础的。该软件可以通过http://www.cis.hut.fi/research/lvq_pak/获得。最早将学习矢量量化引入天文学的是Bazell和Peng，详细的应用可参看他们的文章[1]。

首先将来自可见光和X射线波段的2272个类星体、336个BL Lac天体、1483个活动星系、9967个恒星和484个正常星系的数据作为训练样本和检验样本，我们利用学习矢量量化将样本分成5类，分类结果如表4.1。各类天体的正确率分别为79.1%（类星体）、29.2%（BL Lac天体）、43.8%（活动星系）、85.8%（恒星）、3.1%（正常星系）。对整个样本，在14542个源中有3435个误分类，占23.6%。从该表可以清楚地看到类星体和恒星的分类结果要明显好于其它样本的分类结果。特别指出的是正常星系分类结果最差。大多数的正常星系混入恒星中，而大多数的活动星系混入到类星体中。为方便起见，我们定义活动星系核AGNs包括类星体、BL Lac天体和活动星系，非活动天体non-AGNs包括恒星和正常星系。在这两个波段，类星体和恒星的观测特性比其它天体的观测特性明显，可见整个样本是以点源为主要特征的。所以活动星核通常误分为类星体，non-AGNs通常误分为恒星。150个活动星系误分为恒星，这主要是因为它们具有较弱的X射线辐射。

表4.1　对多类分类问题，来自两个波段的样本的分类结果

分类\已知	类星体	BL Lac天体	活动星系	恒星	正常星系
类星体	1796	126	622	325	59
BL Lac天体	28	98	53	76	14
活动星系	435	106	649	1002	113
恒星	11	5	150	8549	383
正常星系	2	1	9	15	15
正确率	79.1%	29.2%	43.8%	85.8%	3.1%

类似地，将来自三个波段的数据：909个类星体、135个BL Lac天体、612个活动星系、3718个恒星和173个正常星系作为训练样本和检验样本，分类结

果如表 4.2 所示。各类天体的正确率分别为 84.2%、31.1%、60.8%、97.3%、65.9%。对整个样本，5547 个源中 638 个误分占 11.5%。同样该表也表明类星体和恒星的分类结果要明显好于其它种类的分类结果。比较而言，BL Lac 天体、活动星系和正常星系的准确率较低，但是比表 4.1 有明显提高，尤其正常星系的准确率从 3.1%升至 65.9%。但是利用三个波段的数据仍不能将 BL Lac 天体从类星体和活动星系中区分出来，同时活动星系也不能明显地与类星体区分开，但是正常星系却可以很好地与恒星分开。

由表 4.1 尤其是表 4.2，可以看出 AGNs 明显地可与 non-AGNs 分开。我们将样本分为两类活动天体 AGNs 与非活动天体 non-AGNs：来自两个波段的 4091 个 AGNs 和 10451 个 non-AGNs，来自三个波段的 1656 个 AGNs 和 3891 个 non-AGNs。用学习矢量量化对其按两类问题分类。这些样本即为训练样本又为检验样本，分类结果如表 4.3 和表 4.4。准确率对 AGNs 为 92.1%和 97.6%；对 non-AGNs 为 91.6%和 97.1%。对整个样本的准确率达 91.7%与 97.3%。

之后，我们将来自三个波段的样本分成两部分：一部分为训练样本，一部分为检测样本。我们先用 2274 个训练样本训练学习矢量量化，得到分类器，然后用 2773 个检测样本测试该分类器。分类结果列入表 4.5，AGNs 与 non-AGNs 的准确率分别达 96.0%与 96.1%。在 2773 个检测样本中有 2664 个源正确分类，占 96.1%；109 个误分类达 3.9%。

表 4.2　对多类分类问题，来自三个波段的样本的分类结果

分类\已知	类星体	BL Lac 天体	活动星系	恒星	正常星系
类星体	765	48	186	25	1
BL Lac 天体	20	42	19	10	2
活动星系	121	44	372	40	38
恒星	3	0	13	3616	18
正常星系	1	1	22	27	114
正确率	84.2%	31.1%	60.8%	97.3%	65.9%

表 4.3　对两类分类问题，来自两个波段的样本的分类结果

分类\已知	AGN	non-AGN
AGN	3766	880
non-AGN	325	9571
正确率	92.1%	91.6%

表 4.4　对两类分类问题，来自三个波段的样本的分类结果

分类\已知	AGN	non-AGN
AGN	1617	112
non-AGN	39	3778
正确率	97.6%	97.1%

表 4.5　对两类分类问题，来自三个波段的样本分成两组的分类结果

分类\已知	AGN	non-AGN
AGN	795	76
non-AGN	33	1869
正确率	96.0%	96.1%

与支持矢量机的结果类似，我们用不同的样本检验学习矢量量化，分类结果表明来自三个波段的数据的分类结果明显好于来自两个波段的数据的分类结果，不论对两类分类问题还是多类分类问题。可见利用这 10 个参数B-R、B+2.5log(CR)、CR、HR1、HR2、ext、extl、J-H、H-K_s和J+2.5log(CR)将AGN从non-AGN中分离出来完全是可能的。由表 4.5.可知，AGN和non-AGN的分类准确率高达 96.0%以上，即这 10 个特征量作为将活动星系核从非活动星系核中分辨出来是有效的。如果增加其它波段的数据例如射电波段，不仅AGN可以与

non-AGN较好地分开，而且各类AGN之间也可能区分开。其中BL Lac天体和活动星系的准确率较低，可能是样本数太少所致。随着数据在数量和质量上的提高，相信分类结果会更好。

参 考 文 献

[1] Bazell D, & Peng Y, 1998, ApJS, 116, 47

第五章　混合方法

§5.1　主分量分析方法

所有统计学的核心思想就是简化。统计学正是从浩如烟海的数据中提取出简单的人们可理解的事实，从而指导我们的行为。

例如考虑汽车的燃料消费问题。每一个汽车的耗油量取决于它的品牌、使用年限、保养状态以及司机的技术水平。为完全明白一个国家的汽车燃料经济，这需要知道汽车和司机的数量，假如为 10^8。但是要计算出整个国家的油消耗量，10^8的数可用一个量来代替，那就是平均用油量。可见这是何等的简化。

主分量分析方法（Principal Component Analysis，PCA）[1]是一个简化某类特殊数据的工具。设想我们有n个物体，且每个物体有p个参量。例如有n个参加会议的天文学家，我们知道p种情况：他们的身高、体重、发表的论文数、飞的路程和他们的汽车耗油量。这p个参数是怎样相关的呢？是天文学家在机场的逗留时间长的发表论文多呢？还是较瘦的发表的多呢？是汽车不好的坐飞机多呢？还是发表论文多的坐飞机多呢？是否这些相关性仅代表粗糙的相关性？

处理这种问题的传统的方法是画出每两个参量图以寻找相关性。不幸的是，当参数增加时，再这样做显然较复杂，我们很容易陷入参量网的困境中。每一个参数或多或少地与其它参数的混合相关。人们大脑可以轻松自如地处理两三个参数。通过分别画出不同参量对其他参量的图，我们可以了解 5-7 个变量。若超出这个范围，恐怕我们的大脑就要需要帮助了。

主分量分析方法在文献中又称 KL 变换(Karhunen-Loeve transform)或 Hotelling 变换(Hotelling transform)，属于多变量分析方法的一种，正好适合处理这样的问题：当你知道许多事物的多种情况，又想知道这些情况是否彼此相关。主分量分析方法可以找出彼此相关的参量，并把相关的量组合成一个新量，这样大大减小了参量数，同时又不至于损失信息。

§5.1.1　主分量分析方法

对一个样本有 n 个物体、p 个参量 $x_j(j=1,...,p)$，可以找到一组新的正交独立变量，$\xi_1,...,\xi_i,...,\xi_p$，每一个变量是原有变量的线性叠加：

$$\xi_i = a_{i1}x_1 + \cdots + a_{ij}x_j + \cdots a_{ip}x_p$$

确定常数 a_{ij} 使最少数目的新变量可以尽可能地解释样本的变量。那么 ξ_i 称为主

分量。

如果原始数据的大多数变量可以仅用 p 个参量中的几个新变量解释，这样我们就找到了原始数据的较简单的描述，以较少的变量对数据分类。有趣的是主分量分析方法表明原始的数据相关，从而导致新的物理观点。若观测的变量不相关，当然也就不会发现主分量。

主分量分析方法的概念可以通过几何方法或代数方法来介绍。

⑴ 几何方法[1]

考虑 p 个参量的情形，n 个天体为处于 p 维空间的一团云。如果两个或更多的变量相关，云块儿沿着超平面中由这几个相关变量定义的轴的方向分布。当若干变量相关时，较大的扩展产生；或者对少量的变量，相关的变量也较少。主分量分析方法找到了这些扩展方向，把它们当作多维参数空间中的轴。这样每一个天体都可以用新坐标空间中的坐标来表示。该方法首先在原数据空间中通过方差最小化找到最大的扩展（投影）方向。这个方向即为第一主分量，具有最多的天体观测信息。其次，考虑垂直于第一主分量的 p-1 维超平面。在 p-1 维的超平面中，再找出最大的扩展（投影）方向，此为第二主分量。重复该过程，定义出 p 个正交方向。

⑵ 代数方法[2]

考虑一组天体，数目为N（i=1,…,N），具有M个属性(j=1,…,M)。假设r_{ij}为原始的测量量，对数据按如下格式归一化：

$$X_{ij} = r_{ij} - \overline{r_j} \qquad \overline{r_j} = \frac{1}{N}\sum_{i=1}^{N} r_{ij} \tag{5.1.1}$$

其中$\overline{r_j}$为某一属性的平均值。

该数据的标准偏差为：

$$\sigma_j = \sqrt{\frac{1}{N}\sum_{i=1}^{N} X_{ij}^2} \tag{5.1.2}$$

下面给出求解主分量的三种方法：

① 该数据的协变矩阵(covariance matrix)为：

$$C_{jk} = \frac{1}{N}\sum_{i=1}^{N} X_{ij} X_{ik} \qquad 1 \le j \le M \qquad 1 \le k \le M \tag{5.1.3}$$

此矩阵满足的本征方程为：

$$Ce_i = \lambda_i e_i \tag{5.1.4}$$

② 该数据的相关矩阵(correlation matrix)为：

$$R_{jk}=\frac{1}{N}\sum_{i=1}^{N}\frac{X_{ij}X_{ik}}{\sigma_j\sigma_k}\qquad 1\le j\le M\qquad 1\le k\le M\qquad (5.1.5)$$

此矩阵满足的本征方程为：$\mathrm{Re}_i=\lambda_i e_i$　　（5.1.6）

③ 该数据的 SSCP 矩阵(sums of squares & cross products matrix)为：

$$S_{jk}=\sum_{i=1}^{N}X_{ij}X_{ik}\qquad 1\le j\le M\qquad 1\le k\le M\qquad (5.1.7)$$

此矩阵满足的本征方程为：$Se_i=\lambda_i e_i$　　（5.1.8）

这里将本征值按降序排列，即 $\lambda_1>\lambda_2>\cdots>\lambda_M$，则 λ_1 为最大的本征值，表示沿新轴的变量。由 $\frac{\lambda_\alpha}{\sum_\alpha\lambda_\alpha}$ 很容易求出变量的比例。也可以对每个主分量除以 $\sqrt{\lambda_\alpha}$ 重新归一。通过上面的求解过程，我们会注意到主分量分析方法的缺点是它必须假设问题为线性的，而且依赖于变量的归一化方式。

按照拇指规则（a rule-of-thumb），对任何变量大于 1 的主分量应引起特殊重视。同时对于变量大于其它主分量的变量的主分量也应引起重视。主分量为线性分析，所以可以通过结果检验主分量的线性。如果线性分析有效，那么第一主分量(PC1)与第二主分量(PC2)图上应为正常分布，与二者无相关性一致。从数学的角度考虑，不可能存在相关性。但是数据点的非随机分布或者离群数据的存在都会暗示存在非线性或者数据的非均匀性。对于离群数据应排除掉，重新进行分析，或者对坐标轴进行转换如转换成对数坐标，也许该问题就变成了线性问题。主分量分析运用的序列并非实际的测量量，其不仅对非随机分布较敏感，而且对离群数据也较灵敏。这些检验是检测数据中的非线性和发现独立的不同寻常天体的重要工具。

由主分量分析方法的原理，我们不难发现主分量分析方法有如下作用[2]：

⑴ 降维

假设 n 个天体具有 m 个参量，将每一个天体看作一个 m 维矢量。可以找到 $m'<m$ 维主轴充分表示天体的信息，这样省去了分析 $m-m'$ 维参量，将其作为噪音去掉，同时也节约了内存。如果前 m' 个主分量能占到变量的 75%或更多，这时认为降维是可行的。至于临界值具体取多大，由天文学家来定。通常由主分量解释变量的积累的百分比来选择取多少个主分量合适。

⑵ 确定线性混合变量

如果本征值为零，变量在该本征矢的投影也为零，因此该本征矢就成为一点。

另外，若该点为原点，我们会得到 $Xu=\lambda u$，即 $\sum_j u_j x_j = 0$，这里 x_j 为矩阵 X 的第 j 个列矢量。因此，线性混合变量很容易找到。事实上，通过分析二级变量，次级依赖关系可以直接找到。这表示，例如分析三个变量 y_1, y_2, y_3，当输入变量 $y_1^2, y_2^2, y_3^2, y_1y_2, y_1y_3, y_2y_3$，如果下面线性关系存在：

$$y_1 = c_1 y_2^2 + c_2 y_1 y_2$$

我们就可以发现。类似地也可以用对数函数或其它函数。

⑶ 特征提取：选取最有用的变量

特征提取时，我们需要对变量进行简化。首先应找出线性相关的属性，在新的参数空间中高度相关的一些属性在数据处理时应去掉，一些与新的坐标轴接近的变量表明是最相关的、最重要的变量。

⑷ 多维变量的可视化

为便于多维数据的可视化，平面图是必要的。这主要是由于平面图的直观性、可视性和准确性。任何变量的关系都可以通过平面图直观地表示出来。

⑸ 证认隐含变量

主分量分析有时是为了寻找隐含变量。其很容易找到第一主分量和第二主分量，但是对于发现那些相关性较小的轴将有一定的难度。那些在某些轴有较大投影的（尤其具有极值的）数据值得我们进一步研究。这些轴将由若干具有较大的正投影的数据到具有较大的负投影的数据覆盖。

⑹ 数据聚类，或发现离群数据

通过观测平面图，可以发现那些具有相同属性或来自同一过程的数据聚在一起。少量数据偏离这些区域，即为离群数据，从而探测到反常数据或奇怪天体。在某些时候，需要排除离群数据，重新进行数据分析。

§5.1.2 主分量分析方法的应用

主分量分析方法在天文中的应用呈有增无减的趋势，其逐渐渗透到天文学的方方面面。主分量分析方法成为光谱分类的强而有力的工具，如应用于恒星光谱[3,4,5]、类星体光谱[6]、星系光谱[7-14]。可以运用于测光数据[15, 16]、分光光度分析[17]、图像处理[18]、分析线性相关[19]、推导色指数[20]、光度分析[21, 22]、分析星系的金属丰度[23]、研究HI吸收[24]等等。通常由于主分量分析方法具有特征提取和降维的作用，其常常被用于数据压缩或数据预处理，与别的算法结合使用，例如与神经网络结合用于星系光谱分类[25]、与神经网络结合用于天体探测和恒星与星系分

类[26]、与神经网络和傅立叶方法结合用于星系形态分类[27]。

主分量分析方法作为统计方法，可以从大量的观测属性中确定出最小数目的独立的或非相关的变量[28，29]。因此其可以作为数据压缩和分析的工具[3]。Lawrence[30]指出如果我们测得了足够多的观测参数，那么我们可以用主分量分析方法来确定独立的隐含变量的数目。通常在主分量分析方法应用中，主要的目的有两种：（i）揭示参数间的相关性，减小输入空间的维数，因此可以当作一种非监督方法，即数据按自己的方式组织而不需要预先给定种类；（ii）压缩数据从而减小参数的数目，将该结果作为监督的分类方法支持矢量机（SVM）和学习矢量量化(LVQ)的输入。从物理的意义上，数组由若干明显的种类的天体组成，例如恒星、星系和类星体。在这些种类的天体及每一类的子类中，某些物理属性是相关的，其中一些相关性是已知的，而另一些相关性是未知的，这些未知的相关性的发现就其本身而言就是重要的科学结果。物理参量间的相关性的发现意味着减小多维参数空间的维数，这是聚类分析的核心部分。主分量分析方法的优点是当参量间的相关性存在时，大部分线性混合的主要判别能力由前几个本征矢携带，而高阶的本征矢只携带着大部分无用的信息。从而主分量分析方法对全部的参数起到很好的过滤作用，也即其可以降低维数。本文的主要目的：不仅把其作为一种数据压缩技巧，将其结果作为支持矢量机（SVM）和学习矢量量化(LVQ)的输入，而且充当一种非监督的分类方法来探索多维参数空间。

我们对第二章的样本应用主分量分析方法，分析结果列入表 5.1 和表 5.2。表 5.1 给出用主分量分析方法分析所得的本征值及其所占的百分比和积累的百分比，表 5.2 给出主分量分析方法分析所得的本征矢。从表 5.1 可以看出，前 5 个本征矢携带了 54.96%、18.36%、13.03%、5.47%、4.13%的信息或描述能力。前 5 个本征矢的总贡献为 95.95%，加上第 6 个本征矢则总贡献达到 99.32%。这说明作为相当好的近似，前 5 个本征矢实际上已携带了这些参数的大部分信息，其它本征矢的权重可以忽略。表 5.2 给出了前 7 个本征矢，对每一个主分量每一列中的每一个数表示对应输入参量的权重。例如：

$$PC1 = 0.03\times x_1 + 0.06\times x_2 + 0.00\times x_3 + 0.00\times x_4 + 0.00\times x_5$$

$$- 0.01\times x_6 + 0.00\times x_7 + 0.09\times x_8 + 0.33\times x_9 + 0.94\times x_{10}$$

这里的 $x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9, x_{10}$ 分别为B-R、B+2.5log(CR)、CR、HR1、HR2、ext、extl、J-H、H-K_s、J+2.5log(CR)的数值。表 5.2 也表明除了CR、HR1、HR2和extl外的所有参量，都与PC1 相关，PC1 的主要贡献来自J+2.5log(CR)；类似地，除了B-R、CR、HR1 和HR2 外的参量都与PC2 相关，除了B-R和HR2 外的参量都与PC3 相关；PC2 的主要贡献来自J-H和H-K_s，而PC3 的主要贡献来自ext，

B+2.5log(CR)主要对PC6 贡献，B-R主要对PC7 贡献；CR对PC1 和PC2 均无贡献，而HR1 只对PC3 有一点贡献；很显然，HR2 对所有的主分量均无贡献，即HR2 为无效参量。为了更清晰，我们做出了前 3 个主分量的两两图，即PC1 对PC2 图（图 5.1）、PC1 对PC3 图（图 5.2）和PC2 对PC3 图（图 5.3）。尽管恒星、活动星系核和星系有某种程度的重合，但是在新的参数空间中，不同种类的天体占据了不同区域。这几个图描述了主分量分析方法如何将 10 维参数空间压缩为 2 维空间。虽然新的参数空间物理意义不太容易解释，但是显然将这几类天体分开是可能的。

表 5.1　用主分量分析方法分析所得的本征值及其所占的百分比和积累的百分比

主分量 PC	本征值	百分比	积累的百分比
PC1	5145259.50	54.96	54.96
PC2	1718685.88	18.36	73.32
PC3	1219937.50	13.03	86.35
PC4	512328.19	5.47	91.82
PC5	386920.03	4.13	95.95
PC6	315298.69	3.37	99.32
PC7	59770.15	0.64	99.96
PC8	1744.99	0.02	99.98
PC9	1235.49	0.01	99.99
PC10	908.84	0.01	100.00

表 5.2　主分量分析方法分析所得的本征矢

No.	参数	本征矢 1	本征矢 2	本征矢 3	本征矢 4	本征矢 5	本征矢 6	本征矢 7
1	B-R	0.03	0.00	0.00	-0.07	-0.03	-0.19	0.98
2	B+2.5logCR	0.06	0.15	-0.33	-0.42	-0.30	-0.75	-0.19
3	CR	0.00	0.00	-0.01	0.01	0.02	-0.02	-0.02
4	HR1	0.00	0.00	-0.01	0.00	0.00	0.00	0.00
5	HR2	0.00	0.00	0.00	0.00	0.00	0.00	0.00
6	ext	-0.01	0.21	-0.82	0.08	-0.26	0.45	0.09
7	extl	0.00	0.10	-0.37	0.34	0.79	-0.32	-0.01
8	J-H	0.09	0.57	0.21	0.68	-0.35	-0.21	-0.01
9	H-K_s	0.33	-0.75	-0.17	0.45	-0.25	-0.18	-0.02
10	J+2.5logCR	0.94	0.20	0.06	-0.19	0.15	0.14	-0.01

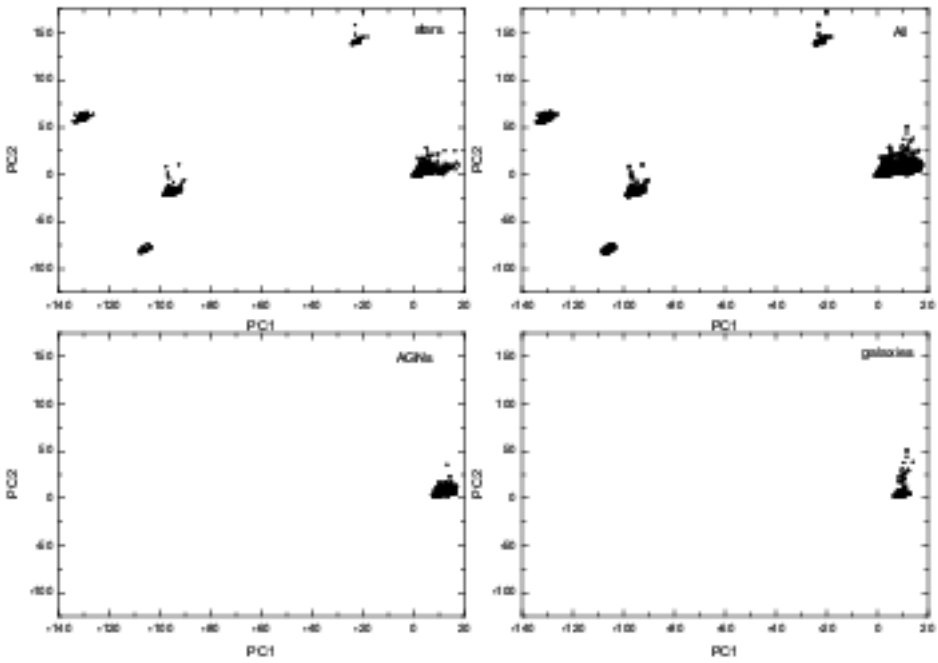


图 5.1　第一主分量对第二分量图

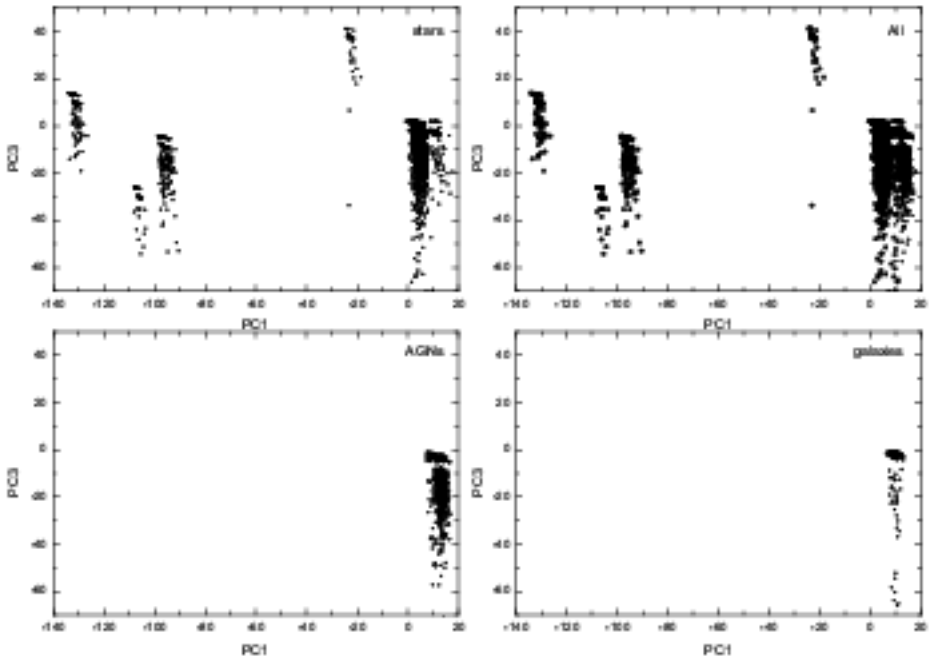


图 5.2　第一主分量对第三分量图

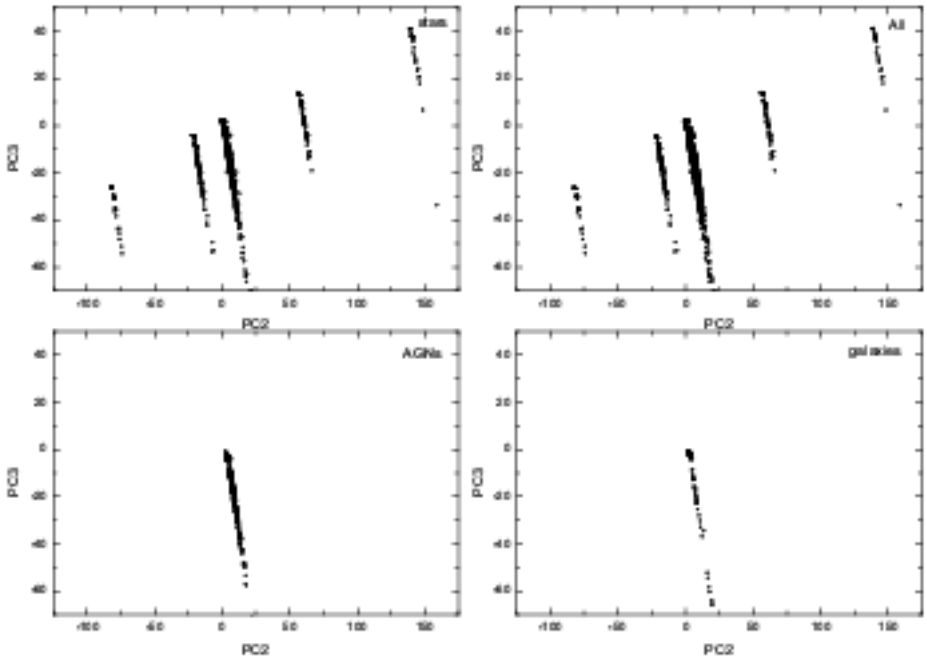


图 5.3　第二主分量对第三分量图

参 考 文 献

[1] Francis P J, “Introduction to Principal Components Analysis” Invited review to appear in ``Quasars and Cosmology", A.S.P. Conference Series 1999. eds. G J Ferland, J A Baldwin, (San Francisco: ASP).

[2] Murtagh F, http://astro.u-strasbg.fr/~fmurtagh/mda-sw/

[3] Murtugh F, & Heck A, 1987, Multivariate Data Analysis. Reidel, Dordrecht

[4] Storrie-Lombardi M C, Irwin M J, von Hippel T, Storrie-Lombardi L J, 1994, Vistas Astron., 38, 331

[5] Bailer-Jones C A L, Irwin M, Gilmore G, von Hippel T, 1997, MNRAS, 292, 157

[6] Francis P J, Hewett P C, Foltz C B, Chaffee F H, 1992, ApJ, 398, 476

[7] Sodre L Jr, Cuevas H, 1994, Vistas Astron., 38, 287

[8] Connolly A J, Szalay A S, Bershady M A, Kinney A L, Calzetti D, 1995, AJ, 110, 1071

[9] Folkes S R, Lahav O, Maddox S J, 1996, MNRAS, 283, 651

[10] Sodre L Jr, Cuevas H, 1997, MNRAS, 287, 137

[11] Galaz G, de Lapparent V, 1998, A&A, 332, 459

[12] Glazebrook K, Offer A R, Deeley K, 1998, ApJ, 492, 98

[13] Ronen S, Aragon-Salamanca A, Lahav O, 1999, MNRAS, 303, 284

[14] Connolly A J, Szalay A S, 1999, AJ, 117, 2052

[15] Buser R, 1976, A&A, 62, 411

[16] Heck A, 1976, A&A, 47, 129

[17] Christian C A, & Janes K A, PASP, 89, 415

[18] Bijaoui A, SAI Library, Algorithms for Image Processing, Nice Observatory, Nice, 1985

[19] Deeming T J, 1968, Vistas in Astronomy, 10, 125

[20] Koorneef J, 1978, A&A, 64, 179

[21] Massa D, 1978, ApJ, 221, 833

[22] Massa D, 1980, ApJ, 85, 1651

[23] Galeotti P, 1981, APSS, 75, 511

[24] Pelat D, 1975, A&A, 40, 285

[25] Folkes S R, Lahav O, Maddox S J, 1996, MNRAS, 283, 651

[26] Andreon S, Gargiulo G, Longo G, Tagliaferri R, & Capuano N, 2000, MNRAS, 319, 700

[27] Odewahn S C, Cohen S H, Windhorst R A, & Philip N S, 2002, ApJ, 568, 539

[28] Kendall M G, 1957, A Course of Multivariate Analysis, Griffin & Co, London

[29] Kendall M G, Stuart A, 1996, in: Advanced Theory of Statistics, Vol. 3, Griffin & Co, London, p. 285

[30] Lawrence A, 1987, PASP 99, 309

§5.2　混合方法

不论数据是已知的或未知的，将数据分成不同的种类，这属于自动分类或聚类的问题。这也是不断增长和快速发展的数据挖掘和数据库中的知识发现的一部分。分类通常分为两类：监督分类和非监督分类。如果数据的种数已知，具有代表性的训练样本可以获得，那么该问题就变为监督分类，常用的监督分类工具如神经网络和决策树。但是比较有趣而无偏差的方法是当数据的种类数未知时，数据需靠自身组织来聚类。非监督分类问题实际上是通过客观的统计方法确定数据中的种类数，然后给出所有天体属于某类的概率。我们以神经网络算法为例，天文学中有许多关于天体分类的工作，例如恒星与星系的分类[1-6]、星系形态的分类[7, 8]、恒星光谱的分类[9-11]。简言之，无论监督分类方法还是非监督分类方法，都各有优缺点。对于非监督分类，它有两点优点：该方法为非监督的，而且不需要对数据进行预处理。但当拥有大而且完备的训练样本时，监督方法将显示出优于非监督方法的优越性。有关监督分类和非监督分类的方法可参看Jain等人的综述文章[12]。

在本文中我们主要介绍两种监督的分类方法：支持矢量机和学习矢量量化，用来将恒星、星系和活动星系核分类。问题是当提供给这两种方法输入时，需要多少个参数和怎样有效地压缩这些参数。当然在保留最少的参数和尽可能多的信息时会有一个界限，我们准备选择一些主分量来研究这些问题。将前 3、4、5 和 6 个主分量(下面间称 3PC、4PC、5PC 和 6PC)分别作为输入，随机地将样本分为两部分，一部分（2774）用于训练，一部分（2773）用于检验。下面将支持矢量机和学习矢量量化这两个方法用于压缩样本，考察主分量数怎样影响分类结果，同时比较这两种方法的分类结果和分类效率。

§5.2.1　主分量分析方法和支持矢量机

首先，我们用包含恒星、活动星系核和星系的训练样本训练支持矢量机，得到支持矢量机的分类器，然后用包含恒星、活动星系核和星系的检测样本检测该分类器，将恒星从活动星系核和星系中分出来，分类结果由表 5.3-5.4 给出。对不同的主分量，总的准确率分别为 98.6%、97.3%、96.9%和 96.6%。

表 5.3　用 3 和 4 个主分量作为支持矢量机的输入的分类结果

	3PC+SVM		4PC+SVM	
分类\已知	AGN 和星系	恒星	AGN 和星系	恒星
AGN 和星系	904	21	885	46
恒星	10	1838	29	1813
准确率	98.9%	98.9%	96.8%	97.5%
总的准确率	98.6%		97.3%	

表 5.4　用 5 和 6 个主分量作为支持矢量机的输入的分类结果

	5PC+SVM		6PC+SVM	
分类\已知	AGN 和星系	恒星	AGN 和星系	恒星
AGN 和星系	884	57	872	35
恒星	30	1802	42	1824
准确率	96.7%	96.9%	95.4%	98.1%
总的准确率	96.9%		96.6%	

其次，用包含活动星系核和星系的训练样本训练支持矢量机，得到支持矢量机的分类器，然后用包含活动星系核和星系的检测样本检测该分类器，将活动星系核和星系分类，分类结果如表 5.5-5.6，对不同的主分量，总的准确率分别为 90.9%，90.4%，90.7%，89.7%。

表 5.5　用 3 和 4 个主分量作为支持矢量机的输入的分类结果

	3PC+SVM		4PC+SVM	
分类\已知	AGN	星系	AGN	星系
AGN	755	10	750	10
星系	73	76	78	76
准确率	91.2%	88.4%	90．6%	88.4%
总的准确率	90.9%		90.4%	

表 5.6　用 5 和 6 个主分量作为支持矢量机的输入的分类结果

	5PC+SVM		6PC+SVM	
分类\已知	AGN	星系	AGN	星系
AGN	754	11	744	10
星系	74	75	84	76
准确率	91.1%	87.2%	89.9%	88.4%
总的准确率	90.7%		89.7%	

§5.2.2　主分量分析方法和学习矢量量化

类似上一节用支持矢量机的做法，我们用训练样本训练学习矢量量化，得到学习矢量量化的分类器，然后用检测样本检测该分类器，首先将恒星从活动星系核和星系中分出来，分类结果由表 5.7-5.8 给出。对不同的主分量，总的准确率分别为 95.6%、96.1%、96.1%和 94.9%。

表 5.7　用 3 和 4 个主分量作为学习矢量量化的输入的分类结果

	3PC+ LVQ		4PC+ LVQ	
分类\已知	AGN 和星系	恒星	AGN 和星系	恒星
AGN 和星系	880	89	864	59
恒星	34	1770	50	1800
准确率	96.3%	95.2%	94.5%	96.8%
总的准确率	95.6%		96.1%	

表 5.8　用 5 和 6 个主分量作为学习矢量量化的输入的分类结果

	5PC+ LVQ		6PC+ LVQ	
分类\已知	AGN 和星系	恒星	AGN 和星系	恒星
AGN 和星系	859	52	842	69
恒星	55	1807	72	1790
准确率	94.0%	97.2%	92.1%	96.3%
总的准确率	96.1%		94.9%	

然后将活动星系核和星系分类，分类结果如表 5.9-5.10，对不同的主分量，总的准确率分别为 89.8%，90.2%，90.6%，90.3%。

表 5.9　用 3 和 4 个主分量作为学习矢量量化的输入的分类结果

	3PC+ LVQ		4PC+ LVQ	
分类\已知	AGN	星系	AGN	星系
AGN	741	6	745	7
星系	87	80	83	79
准确率	89.5%	93.0%	90.0%	91.9%
总的准确率	89.8%		90.2%	

表 5.10　用 5 和 6 个主分量作为学习矢量量化的输入的分类结果

	5PC+ LVQ		6PC+ LVQ	
分类\已知	AGN	星系	AGN	星系
AGN	749	7	745	6
星系	79	79	83	80
准确率	90.5%	91.9%	90.0%	93.0%
总的准确率	90.6%		90.3%	

参 考 文 献

[1] Odewahn S C, Stockwell E B, Pennington R L, et al. 1992, AJ, 103, 318

[2] Naim A, Lahav O, Sodre L Jr, et al. 1995, MNRAS, 275, 567

[3] Mähönen, P. H., & Hakala, P. J. 1995, ApJ, 452, L77

[4] Miller A S, & Coe M J, 1996, MNRAS, 279, 293

[5] Bertin E, Arnout S, 1996, AAS, 117, 393

[6] Bazell D, & Peng Y, 1998, ApJS, 116, 47

[7] Storrie-Lombardi M C, Lahav O, Sodre L Jr, et al. 1992, MNRAS, 259, 8

[8] Lahav O, Naim A, Sodre L Jr, et al. 1996, MNRAS, 283, 207

[9] Bailer-Jones C A L, Irwin M, von Hippel T, 1998, MNRAS, 298, 361

[10] Allende Prieto C, Rebolo R, Lopez R J G, et al. 2000, AJ, 120, 1516

[11] Weaver W B, 2000, ApJ, 541, 298

[12] Jain A K, Duin R P W, Mao Jianchang, 2000, Pattern Analysis and Machine Intelligence, 22, 4

§5.2.3 小结

对来自三个波段的数据进行主分量分析，考察各个参数对分类的影响。由主分量分析结果可知，参数CR、HR1 和HR2 对天体的分类几乎不起作用，这主要是由于参数CR与距离有关而不同波段的流量比与距离无直接关系，HR1 和HR2 的误差较大以至于各种天体重叠。参数J+2.5log(CR)、J-H、H-K_s和ext对第一、第二和第三主分量贡献较大；B+2.5log(CR)在当主分量数大于 2 时贡献作用加大，B-R则在主分量数大于 6 时作用才加大。这可能由于相对于红外J、H和K_s星等，B星等和R星等的误差远大的多。因此，那些对分类有用的参数不仅与其定义有关而且与它们的观测误差有关。只有那些与流量比例有关的参数才对分类起决定性作用，从天体物理学的角度来看，参数的选择标准是建立在各类天体的不同的能谱分布基础之上的。

由统计结果可以看出，这些算法可以很好地将恒星从河外天体中挑选出来，并且这些算法可以高效地将真正的恒星分为恒星和非恒星分为非恒星，即误分的概率相当小。对比 PCA+SVM 和 PCA+LVQ 方法的分类结果，我们发现 6PC+SVM 方法的最低准确率（96.6%）也比 4\5PC+LVQ 方法的最高概率（96.1%）要高，不过它们的准确率都大于 94%。在活动星系核与正常星系分类时，PCA+SVM 和 PCA+LVQ 两种方法分类效果相当，都约为 90%。对支持矢量机而言，似乎要提高准确率最好使用 PCA 降维；对学习矢量量化而言，是否降维对分类准确率的影响不大。为了检测分类器的效率，我们用整个样本来训练，并记下它们所用的 CPU 时间，结果列入表 5.11 中。由表 5.11 可得出，对支持矢量机而言，主分量数越多，所用的时间越少；对学习矢量量化而言正好相反，主分量数越多，训练的速度越慢。因此，若提高速度，当使用支持矢量机时不需要降维，但使用学习矢量量化时则需要降维。但是 PCA+LVQ 方法的最慢速度 6 秒远远快于 PCA+SVM 方法的最快速度 56.9 秒。因而，如果只考虑准确率时，我们最好选用 3PC+SVM，但是对较大的样本考虑到效率，最好选用 3PC+LVQ。另外，活动星系核和正常星系的分类准确性不如恒星的高，这可能主要由于它们的样本数小于恒星的（1829 与 3891）。一旦拥有与恒星数目相当的星系和活动星系核，我们可以用这些的方法很容易地将这三类天体分开。

表 5.11 对不同的主分量，PCA+SVM 和 PCA+LVQ 方法所占用的 CPU 时间

方法	CPU 时间\输入	3PC	4PC	5PC	6PC
SVM	CPU 时间（秒）	94.3	76.4	63.5	56.9
LVQ	CPU 时间（秒）	2	2	3	6

我们提出了两种算法及其与主分量分析方法的混合方法运用于天体的多参数分类。这些算法的有效性是基于训练样本的，并通过检测样本来检测其可靠性。因此，这些样本尤其是训练样本越完备越准确越好。这些方法正是在高效地提取来自多种巡天得到的多波段数据中的信息的需求下引入的。面对多维参数空间，我们可以先用主分量分析方法得到有用的参数和降低维数。然后将 PCA 的结果作为其它算法的输入。

对比计算结果，我们得出结论 SVM/PCA+SVM 和 LVQ/PCA+LVQ 方法是有效的分类多波段数据的算法，而且 SVM/PCA+SVM 表现出优于 LVQ/PCA+LVQ 的分类效果。在若干情形下，这两种方法给出的结果相当。通常 SVM/PCA+SVM 的分类正确率高，而 LVQ/PCA+LVQ 在计算速度上要快得多。统计结果表明 SVM 与 LVQ 相当甚至优于后者，这主要是由于二者所用的原理不同。支持矢量机的算法体现了结构风险最小化（SRM）的原理，这明显优于传统的神经网络所用的经验风险最小化(ERM)的原理。SRM 最小化预期风险的上限，而 ERM 只是最小化训练样本的误差。因而大多数神经网络找到的超平面不一定是最优的。实际的操作中，许多神经网络只是随机移动一条线，直到所有的训练样本留在线的一边。这样不可避免使得训练数据点较接近分界线，从而不是最优的分界线。相比较而言，支持矢量机通过训练可以得到具有最大分类边界的分类器，也即找到了最优的分界线。正是基于这点不同，使得支持矢量机具有较强的推广能力。

这些方法适合多种种类的巡天数据的应用，如多色巡天、测光巡天和光谱巡天，但主要是为了获得大的样本用以统计研究。当然，这些方法只是给出了预选样本，要想确认一颗天体具体属于某类，还需要通过观测证认。从巡天数据中挑选出一批候选体，可以减小时间和精力的浪费。

这些方法能够对天体的分类有效地起作用，正是利用了天体的基本特征如流量比J+2.5log(CR)、J-H和H-K_s。由于数据的误差可造成结果的歪曲并减小信息的输出，因而源的分类不仅依赖于数据的数量和质量，还依赖于算法的原理。既然高分辨的多波段数据持续增长，我们相信随着数据的完备和数据的数量与质量的提高，分类结果会越来越好。可以用获得的分类器对未证认的RASS源进行分类，从而用于统计分析，或探索一些非监督的分类算法或离群数据发现算法来发现某些特殊的、稀有的、不同寻常的、甚至完全新种类的天体或天文现象。简言之，随着虚拟天文台的发展，这些方法有助于发展国际虚拟天文台的工具箱。

第六章　结论和展望

本论文从多波段天文学的兴起谈起，综述了虚拟天文台产生和发展的必然性、结构和科学目标、以及其对天文学发展所起的巨大的推动作用。并且指出数据挖掘技术在虚拟天文台中的成功应用，是虚拟天文台充分发挥作用的关键所在。着重阐述了数据挖掘和知识发现兴起的必然性、近几年的发展、过程、任务和方法，以及其所面临的问题。这也正是天文学中的数据挖掘和知识发现得以发展和完善的外在动力。结合天文学的特点和要求，描述了天文学中的数据挖掘和知识发现的相关知识。并就天文学家感兴趣的问题——发现不同寻常的、稀有的或新类型的天体或新天文现象，介绍了一些离群数据的发现和计算的方法。

第二章介绍了活动星系核的特征和分类，并且着重描绘了各种预选源方法的优缺点及其所达到的准确度。然后介绍了多波段交叉证认的原理和方法。我们以ROSAT 亮源表与弱源表、USNO-A2.0、2MASS、Veron (2000)、SIMBAD 和 RC3 星表为例，介绍了交叉证认获得样本的过程。由多波段数据发现各类天体在多维参数空间中表现特征的不同，从而其作为自动化分类方法的样本是合理的、可行的。

第三章至第五章我们提出了两种方案用来研究天体在多维参数空间中的分布。第一种方案：利用多波段数据，用自动的分类方法支持矢量机（SVM）和学习矢量量化（LVQ）对天体分类，对比了采用两个波段数据与三个波段数据的分类结果，发现随着波段的增加，分类效果越好。可见提取的天体信息越多，越有利于天体分类。第二种方案：针对未来天文数据维数过高的问题，我们探索了这两种方法与主分量分析方法（PCA）的混合方法，即 PCA+SVM 和 PCA+LVQ。

在天文学的巡天观测中，最明显的天体种类有恒星、星系、类星体和奇异天体。这些天体根据它们的物理特性如形态或能谱分布可以进一步分为若干子类。因此，形态、颜色或特殊的谱特征可以作为分类标准用于巡天数据的分类。考虑到形态，那些点源通常被认为是恒星或类星体，星系通常为展源。所以只要将恒星和类星体分开，不管星系是混在恒星中还是类星体中，通过形态很容易将其挑出。因而我们提出的方法可用以预选类星体候选体。

通过前面的分析，建立在天体的多波段特性基础上，我们提出的自动分类方法：支持矢量机和学习矢量量化及它们与主分量分析方法的混合方法是合理的、可行的。它们在天文中会有如下应用：

⑴ 通过这些方法获得的分类器可以用于从大的巡天中选取活动星系核候选体，从而避免没必要的时间和精力的浪费。例如：多于 65%的RASS(the ROSAT All-Sky Survey)[1, 2]源仍未证认[3]。因此，这些方法可以有效地从RASS巡天中选取活动星系核候选体。从而可以得到完备的活动星系核样本，用以研究宇宙学和

大尺度结构。

⑵ 这些方法可以用于对各种天体进行分类，例如恒星分成不同光谱型、星系按照哈勃序列或形态分类、活动星系核分成类星体、BL Lac天体和活动星系。不只用多波段数据，还可以用光谱数据、测光数据，也可以增加波段选取更多的参数。我们发现确实波段越多分类效率越高[4]。利用各种已知天体的光谱数据训练支持矢量机和学习矢量量化，我们可以将获得的分类器用以大型的光谱巡天，如中国正在建设中的LAMOST望远镜。

⑶ 它们可以有效地处理高维数据，如 10 维或更高维。在独立的或联合的巡天数据中，典型的数组含有约 10^8-10^9个源，对每个源约有 100 个测量属性，即约 10^9个数据矢量分布于 100 维参数空间。主分量分析方法可以作为一种预处理技巧自动地找出那些含有大部分信息的维数。换句话说，其不仅能找到相关的参量，而且有助于降维。这样分类算法可以在较低维的空间中实施，从而降低计算的复杂度。但是主分量分析方法是一种非监督的线性特征提取方法，不适合非线性的情况。当面对非线性的情况，应当使用非线性特征提取技巧，例如带核的主分量分析方法(Kernel PCA)和多维尺度分析（multidimensional scaling）[5]。

⑷ 为提高效率，我们提出了两种混合算法PCA+SVM和PCA+LVQ。类似的工作还有：Folkes等人[6]和Lahav等人[7]应用主分量分析方法和人工神经网络（PCA+ANN）对星系的光谱分类；Andreon等人[8]用PCA压缩图像数据，而后用ANN对恒星和星系分类。Odewahn等人[9]先用傅立叶方法处理数据，然后用PCA+ANN对星系从形态上进行分类。很显然，当面对丰富的数据环境，如虚拟天文台中多个巡天数据的融合提出的各种各样的问题，一些算法和模型的互操作和重复利用是必要的，也是必需的。

⑸ 利用交叉证认得到的多波段数据，我们可以用一些非监督算法或离群数据发现的算法，来发现某些特殊的、稀有的、不同寻常的、甚至完全新种类的天体或天文现象。

本文工作的主要目的是为了探索一些自动分类算法在多波段数据处理中的应用。用支持矢量机(SVM)和学习矢量量化方法(LVQ)及这两种方法与主分量分析方法(PCA+SVM/LVQ)的混合方法，对多波段数据进行分类。随着数据的更新和完备，可以重复该工作以得到更加优越的分类器。也可以再运用一些其它数据表如 DPOSS 和 SLOAN 可见光数据表、FIRST 和 NVSS 射电数据表及其它波段的数据表，我们将处理一些只有用多波段数据才能解决的问题如：① 特征化近邻星系的多波段形态；② 用多波段的方法解释活动星系和类星体；③ 证认活动的天体系统与其环境的关系。还可以将该方法用以其它种类的数据（测光、光谱、图像等数据）或这些种类数据的混合数据的分类。另外，也可从方法的研究着手，

探索适合天文学的分类、聚类和离群数据的探索方法等技术。具体而言，将来工作的方向可以有如下几种：

⑴ 数据样本不变，将其它的数据挖掘技术运用于该样本；

⑵ 数据样本更新，但数据挖掘方法不变；

⑶ 数据样本更新，并利用其它的数据挖掘技术进行挖掘；

⑷ 改变数据类型，选用光谱数据、图像数据、测光数据、其他类型的数据或几种类型数据的混合数据，运用合适的数据挖掘技术进行挖掘；

⑸ 计算交叉证认的概率，可以处理一对多、多对一等情况的源，挑出可靠性较高的源为对应体，从而进行数据挖掘；

⑹ 利用前任何一步获得的分类器为大的巡天项目（如 LAMOST）预选源；

⑺ 利用前 5 步的任何一步获得的分类器对大型的样本分类，用以统计研究；

⑻ 发展一套适合天文数据特点的离群数据探测方法；

⑼ 发展自动的方法确认交叉证认的对应体；

⑽ 从长远的角度，应发展自动的、高效的、标准化的天文数据挖掘技术，从而成为虚拟天文台工具箱的一部分。

随着地面和空间的大面积的多波段巡天数据的获得，天文数据的体积和质量仍在飞速增长。各种巡天项目将得到数以千亿字节甚至万亿字节的天体从射电波段到 X 射线波段的形态和光谱信息。即天文学正在步入一个全新的时代——“数据雪崩”时代。所获得的数据就其本身而言是十分突出的，即高分辨率的、多波段的各种数据如测光、光谱和图像等数据。由数据增长而带来的数据查询和获得问题，将随着国际虚拟天文台的建立而得到解决。海量数据带给我们的不仅有挑战而更多的是机遇。尤为重要的是将这些丰富的数据联合起来考虑，我们将会对不同种类的天体产生突破性的理解。多波段信息的融合将提供更加准确的天体性质的描述，以及一些测量量如能谱分布、形态等，这将对天体的物理性质给予限制。当然对于多波段数据，革命性的步骤在于如何将这些数据融合起来。要想融合在一起，就需要利用各种数据的共同属性——位置。但是这种方法隐藏着一些缺点，尤其在处理不同的分辨率和观测深度的多波段数据时。因此需要探索一种比较新的方法，即利用获得的数据又运用天文理论，最优化地运用概率方法确定源的相关性。多波段数据分析流程图如图 6.1 所示。由该图可知：通过交叉证认，我们可以获得多种情况下（一对一、一对多、多对一、一对无和多对无）的多波段数据。对前三种情况下的数据进行概率分析，挑出高概率的数据，得到高质量的各种天体类型的数据表，可以用于具体研究如数据挖掘（分类分析、聚类分析、关联规则分析、序列模式分析、偏差分析、依赖关系分析，等），或者由光谱证认发现已知类型的新天体。对于低概率的数据可以用于统计研究及其后继的研

究，例如：光谱观测许多低质量的恒星来发现冕发射的特性；研究天体的类型与 X 射线发射的关系，如脉动变星的 X 射线发射。对一对无和多对无的情况，我们要引起注意，加以特别分析，也许会发现一些稀奇古怪的天体。

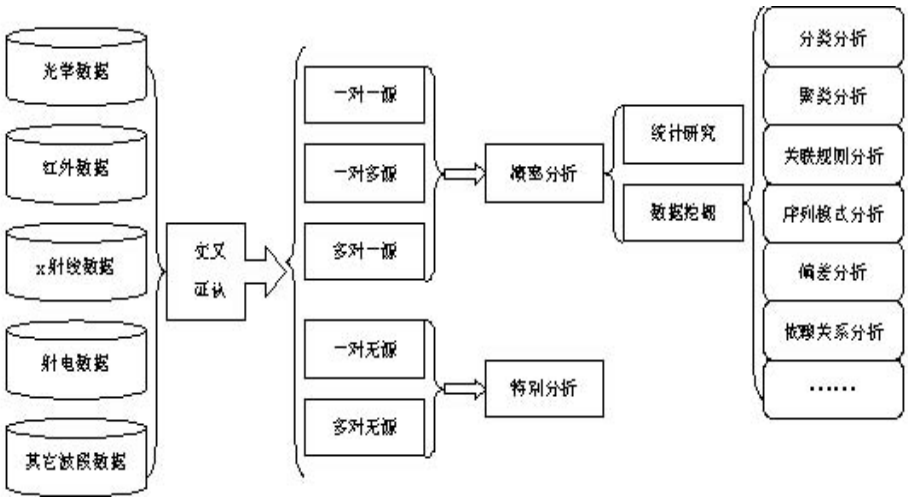


图 6.1　多波段数据分析流程图

从正在增长的海量天文数据中最大化知识提取的关键因素在于成功地应用数据挖掘和知识发现技术。这将为发展下一代的科学分析工具打下基础，从而可以重新定义科学家从数据中发现知识的方法和工作方式。从第一章的数据挖掘和知识发现的综述中，我们已了解到数据挖掘技术已经渗透到各种数据丰富的领域，并且取得了很大的成功。在天文中我们可以将数据挖掘技术应用到大型数据的数据挖掘中。例如：

⑴ 种类繁多的分类技术，包括各种决策树、近邻规则分类器、神经网络、以及本文中提出 SVM 和 LVQ 等用于对天体进行分类或聚类。研究客观上聚为一类的数据是否对应物理上有意义的明显种类的天体？是否发现了新的天体类型或新的天体？是否可用客观的方法对某类天体进行细分类（如哈勃分类、恒星光谱型）？

⑵ 聚类方法，如预期最大方法（the expectation maximization，EM）用混合模型可以发现数据中有意义的类别，给出描述性的结论，并且可以对数据进行密度估计。从统计学和客观的角度，探索数据中究竟有多少种类的天体。这是一种有效的方法将数据聚类，用于具体的研究，例如某些用户只用恒星，其他用户只用星系，或者只用红外超的天体，等。

⑶ 用遗传算法发展更高级的探测和监督的分类方法。这将是特别有用的技术在图像和星表数据共同作用的领域。

⑷ 发展聚类算法和离群数据探测方法，用以探测稀有的、反常的或不同寻

常的天体，例如挑选出参数空间中的离群数据用以进一步研究。这不仅包括已知但稀有的天体，例如褐矮星、高红移类星体，而且包括可能新的、以前未知类型的天体和天文现象。

⑸ 用半自动的人工智能和软件工具研究大的参数空间，探索不同寻常的事件和天体类型出现的频率。采用这样的方法考察数据怎样组织才能最优化地探索参数空间。

⑹ 发展高效的、新的数据可视化和表达技术，这样可以表达大部分多维信息，以直观和清晰的表达方式让用户一目了然。用三维图加上天体的形状和颜色来表达多维参数空间的信息，并且交叉应用图像和星表数据。

上面的事例远远超出仅提供处理大型数据的帮助。这些软件工具应具备独立的或协作的科学发现能力。它们的应用可以大大地提高科学家的科学产出和创造力。在各种计算机科学和统计学领域中适合上述任务的许多高级工具已经存在或正在发展中，有些可以直接移植过来，有些则需要稍微调整即可用于天文研究。在创建虚拟天文台过程中，最重要的科学要求之一是架起学科之间的桥梁，将现代的数据管理和分析软件技术引入天文学和天体物理的科学研究中。

天文中的大部分的巡天工作和仪器的发展都是源于研究星系的形成和演化。该工作的转折点在于在各种各样的多波段数据的基础上或者通过虚拟天文台，估计遥远的星系样本的各种物理参数。目前这些参数的估计仅仅用了相对有限的测量量。例如遥远星系的形态类型是通过可见光的图像来分类的，而红移则是通过它们的宽波段能谱分布确定的。对于那些相对小的样本（主要是那些 Hubble 深视场中的样本），我们既可以确定它们的形态也可以确定它们的红移。通常天体的能谱分布与其形态是紧密相关的，因此能谱分布可以为天体形态的确定提供参考，同时观测的形态数据可以对红移给予限制。在不久的将来，将有远比 Hubble 深视场中的样本多得多的数据同那些与之相关的多个波段的数据一起获得。尽管某些数据只是有限的多波段数据，但是建立在比较完备的样本基础上的先验知识可以对这些有限样本的参数给出较准确的估计。

用丰富的异构数据确定参数的技术有神经网络、主分量分析方法等数据挖掘和知识发现技术。目前这些技术在天文中只有有限的应用。例如：主分量分析方法用以确定类星体光谱的主要元素；神经网络对恒星的光谱分类，通过 Hubble 深视场中遥远星系的图像确定其形态；星系的合成光谱与观测的标准能谱分布相比来估计测光红移。在许多的多波段数据可以在线获得时，我们希望创建一个系统，其可以通过训练样本的学习，然后确定尽可能多的物理参数。从较简单的层次讲，用光谱红移和多波段的测光数据组成训练样本，然后利用从训练样本中获得的知识，可以从标准能谱中确定出红移值。从较复杂的角度讲，利用多波段的

形态数据、光谱数据和测光数据作为训练样本，可以估计遥远星系的形态类型、星族的混合成分、恒星形成率以及恒星、气体和暗物质所占质量的百分比。

当然，建立在广泛的信息基础上的物理参数估计问题是一个重要的问题，不仅仅限于天文研究。任何在天文中对该问题的解决方案可以直接应用于其他领域的纯理论和纯应用的研究中。例如：通过确定新的信息是否由特殊的仪器或仪器的特殊结构产生的，来确定具体的输出参数的可靠性，这样的系统可以用于仪器或试验的设计中。在非参数和模型独立的情况下，探索仪器或试验的设计的各种边界值对参数估计的影响。

不论是用于天文领域或别的领域的数据挖掘和知识发现技术的发展和软件系统的应用，我们的最终目的是要发展一套适合虚拟天文台的各种用途的系统。利用所有可获得的数据，无论是原始的还是加工的多波段的关于某天区或天体的信息，可以用于估计天体的类型、红移、物理参数等等，或者发现不同寻常的天体。虚拟天文台的全方位的应用仍处于探索阶段，以期将一些算法和软件模型化做成自动化的系统，来处理来自各种巡天的不同类型的多波段数据。一些算法输出的天体的参数列表要格式化成虚拟天文台要求的格式，以作为输入数组去估计物理参数。这项工作的天文应用将利用比以前高得多的效率和质量的数据，来发现巡天中不同寻常的、稀有的或完全新类型的天体或新天文现象和估计大尺度结构的参数。从而为天文学理论的发展和完善奠定坚实的基础，同时也为其他学科的发展和完善提供重要的参考。相信随着虚拟天文台的建立和使用，数据挖掘和知识发现技术的成熟和应用，天文学将绽放出更加鲜丽的花朵。天文学家应顺潮流而动，把自己培养成适合时代需要的全方面的人才，在天文学发展的历史长廊中绘出自己的大手笔。

参考文献

[1] Voges W, 1992, in Proceedings of the ISY Conference “Space Sciences”, ESA ISY-3, ESA Publication, p.9

[2] Voges W, 1997, The All-Sky Survey and Pointing Catalogues of ROSAT. In: Di Gesu V, Duff M J B, Heck A, et al.(eds.) Data Analysis in Astronomy V. World Sci. Publ., Singapore, p.189

[3] Voges W, Boller Th, Dennerl K, et al., 1996, MPE-Report 263, 637

[4] Zhang Y-X, Cui C-Z, Zhao Y-H, 2002, in Jean-Luc Starck & Fionn D Murtagh (eds.), Astronomical Data Analysis II, Proc. of SPIE, 4847, p.371

[5] Jain A K, Duin R P W, Mao Jianchang, 2000, Pattern Analysis and Machine Intelligence, 22, 4

[6] Folkes S R, Lahav O, & Maddox S J, 1996, MNRAS, 283, 651

[7] Lahav O, Naim A, Sodre L Jr, et al. 1996, MNRAS, 283, 207

[8] Andreon S, Gargiulo G, Longo G,.2000, MNRAS, 319, 700

[9] Odewahn S C, Cohen S H, Windhorst R A, et al. 2002, ApJ, 568, 539

致　谢

光阴似箭、日月如梭，转眼三年过去了，不能忘怀的是自己的恩师赵永恒研究员这三年来对我的悉心关心和照顾，尤其在本论文从选题到最后的完成过程中所给予的指导和帮助。回首往事，历历在目，忘不了赵老师的忙碌的身影、广博的学识、兢兢业业的工作、谦虚诚恳的为人、博学多才的能力。他常常严于律己、宽以待人、克己奉公、脚踏实地。他的思维敏捷，经常在我迷惑不解时，给予热情的指点，使我顿开茅塞、拨云见日，有时真有“山重水复疑无路，柳暗花明又一村”的感觉。他的胸襟宽广，无论发生什么事，从未见他动过容，对我们学生从不指手画脚，希望我们自己能够辨明是非，充分发挥我们的能动性。在一个人的成长过程中能遇见这么好的恩师指点迷津，是一个人的荣幸。我在这里向他致以崇高的由衷的谢意。在今后的人生旅程中，我一定会铭记赵老师的敦敦教诲，以赵老师为学习楷模，踏踏实实地做好自己的工作，不辜负老师的厚望。

感谢我的硕士导师张波教授在这三年来对我的殷切关心、不断的鼓励、耐心的指导和帮助。感谢何连发教授对我的工作的帮助和支持，坦诚的话语常萦绕于胸，慈父般的神态常浮现于脑海。

感谢国家天文台的胡景耀老师、赵刚研究员、邓李才研究员、武向平研究员、魏建彦研究员、周旭研究员、彭勃研究员和韩金林研究员，虽与他们交往不多，但他们的宽容谦虚的仁者风范和严谨的治学态度深深地影响着我。

感谢杜红荣老师对我的热情关怀和帮助。感谢梁艳春博士的无微不至的关切、问候、支持和鼓励。感谢郑宪忠博士的热心帮助和鼓励。感谢陆烨博士的有意义的建议和帮助。感谢石火明博士的朴实耐心的忠告和帮助。感谢施建荣博士的热心的问候和鼓励。

感谢国家天文台 LAMOST 项目总部的老师：苏洪钧总经理、王刚研究员、陈英老师、袁晖老师、孙盛慈老师、李有刚老师、李颀老师、冯磊师傅和门力老师的关心、照顾和帮助。

感谢国家天文台 LAMOST 全体成员的关心和帮助。三年来形成的兄妹情、姐弟情、姊妹情，使我终生难忘。忘不了一起嬉笑打闹的情景，忘不了互帮互助互学的场景，忘不了……。谦虚大度的罗阿理、诚恳随和的陈建军、聪明能干的崔辰州、谨慎稳重的程林鹏、学识渊博的王伟、热情开朗的卞维豪、谦虚谨慎的覃冬梅、认真踏实的赵瑞珍、豁达朴实的张昊彤、坦诚直率的桑健、憨厚幽默的

吴潮、聪慧活泼的邵惠娟、聪明灵活的朱光华、敏于好学的许馨、机智爽快的刘中田，温厚谦和的李博，风趣幽默的贾磊，一个个在我脑海中活灵活现。在此向大家表示我诚挚的谢意，感谢大家对我的热情帮助和殷切鼓励。

感谢国家天文台好友苏彦、李冀、王菲鹿、杜翠花、孔民芝、刘继红、边霞和所有其他关心和帮助过我的人，在此一并表示衷心的感谢。

尤其感谢我的父母和我的爱人对我的学习的理解和支持。

话语千千万，一切源于一个“爱”字！

在此谨祝：好人一生平安！

发表论文目录

1. **Zhang, Yanxia**, & Zhao, Yongheng, " Classification in the Multidimensional Parameter Space: Methods and Examples ", 2003, PASP, 115, 1006
2. **Zhang, Yanxia**, & Zhao, Yongheng, “Automated Clustering Algorithms for Classification of Astronomical Objects”, 2003, submitted to A&A
3. **Zhang, Yan-xia,** & Zhao, Yong-heng, “Learning Vector Quantization for Astronomical Objects Classification”, 2003, ChJAA, Vol.3, No.2, 183
4. **Zhang, Yanxia**, Cui, Chenzhou, Zhao, Yongheng, “Classification of AGNs from stars and normal galaxies by support vector machines”, 2002, in: Jean-Luc Starck & Fionn D. Murtagh (eds.), Astronomical Data Analysis II, Proc. of SPIE, 4847, 371-178
5. **Zhang, Yan-xia,** Zhao, Yong-heng, Cui, Chenzhou, "Data Mining and Knowledge Discovery in Database of Astronomy", 2002, PABei, 20 (4), 312
6. **Zhang, Y.-X.,** Zhang, B., Li, J., et al.,” The average abundance of heavy elements in metal-poor stars in different metallicity ranges”, 2001, ChA&A, 25, 187
7. Zhao, Yongheng, **Zhang, Yanxia**, Cui, Chenzhou, ”Classification of Active Objects in the Multiwavelength Parametric Space”, 2002, VO meeting in Germany, in press
8. Zhang, Bo, **Zhang, Yan-xia,** Li, Ji, Peng, Qiu-he, “A Statistical Model for Predicting the Average Abundance Patterns of the Heavier Elements in Metal-Poor Stars”, 2002, ChJAA, 2, 429
9. Zhang, Bo, **Zhang, Yan-Xia**, Liu, Jun-Hong, et al., ”The abundance distribution of neutron-capture elements in metal-poor stars”, 2001, ChA&A, 25, 298
10. Zhang, B., **Zhang, Y. X.**, Liu, J. H., et al., “The abundance distribution of elements captured by neutrons in metal-poor stars”, 2001, AcASn, 42, 22
11. Zhang, B., Li, J., **Zhang, Y. X.**, et al., ”The abundances of neutron-capture elements in metal-poor stars”, 2000, PABei, 18, 238
12. Zhang, Bo, Li, Ji, Wang, Yue-Xiang, **Zhang, Yan-Xia**, et al., ”Two R-Process Components in Ultra-Metal-Poor Stars: The Neutron-Capture Element Distribution of CS 22892-052”, 2002, Ap&SS, 280, 325
13. Cui, Chenzhou, Zhao, Yongheng, Zhao, Gang, **Zhang, Yan-xia,** "Technical Progresses of International Virtual Observatories", 2002, PABei, 20 (4), 302