分类号 _____ 密 级 _____

UDC _____ 单位代码 10094

河北师范大学

硕士学位论文

天文中的分类和回归方法初探

李 丽 丽

指 导 教 师： 杨大卫 教授 河北师范大学

张彦霞 副研究员 中科院国家天文台

赵永恒 研究员 中科院国家天文台

专 业 名 称： 理论物理 研 究 方 向： 天体物理

申请学位级别： 硕 士 答 辩 日 期： 2007 年 6 月

中 文 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，使得天文数据急剧增长。面对海量天文数据的冲击，借助数据挖掘技术来探索隐藏在数据中的有用信息势在必行。在此背景下，本文利用了一些适合于天文数据特点的数据挖掘技术和方法，试图完成对天体的自动分类和预测测光红移任务。主要工作包括以下三部分：

（1）基于神经网络预测测光红移

测光红移有其自身发展的优势，尤其对大量无法探测到光谱的暗弱天体而言，更具有一定的统计意义，它能使我们更好地了解宇宙大尺度结构和星系的形成与演化。本文利用了数据挖掘中的神经网络方法对SDSS巡天数据中的星系进行测光红移的测量。主要使用了三种有效的星等参数：*Petrosian* 星等、模型（*model*）星等和红化校正（*deredening*）星等，进行了三组实验，并且每组实验又考虑了星系颜色和各波带处50%和90%的*Petrosian*流量半径，通过实验对比得出对神经网络测红移而言最好的一组参数组合，为天文学家选择参数提供参考。

（2）*K*近邻方法对天体自动分类

K 近邻方法在数据挖掘中是一种简单而又灵活有效的分类方法，已被广泛应用到许多领域中。本章使用 *K* 近邻方法对天体自动分类，根据各类天体在多个波段的不同表现性质，将活动星系核从恒星和星系样本中分离出来。使用的样本是由多个数据库交叉认证获得的多波段数据，参数包含天体不同波段的相关信息。用 *K* 近邻方法完成自动分类任务，分类精度达到了97.73%，说明 *K* 近邻方法作为自动化分类方法是比较合理的、可行的。

（3）星系形态分类及测光红移

我们就如何提高测光红移的精度提出了一种新方案：先用 *k-means* 自动聚类算法将星系样本按形态聚为两类：早型星系和晚型星系，再分别使用神经网络方法预测这两类样本的测光红移，计算出它们的结合精度。通过实验比较分类后得出测红移效果与直接使用星系总样本预测红移的效果，发现先分类再分别对各类预测红移给出的结合精度要高于直接使用总样本预测红移的精度，而且最好的剩余标准偏差为 0.0192，尤其对早型星系的剩余标准偏差竟可以达到 0.0164。研究结果表明我们提出的方案是合理的、可行的、有效的。

关键词：数据挖掘，神经网络，自动分类，测光红移

ABSTRACT

With the establishment of ground-based and space-based observation stations, the improvement in detector sensitivity, the increase in the telescopic caliber and the rapid development of network technology, the astronomical data increase sharply. Faced on the impact of the huge data, it is necessary to explore the potential and useful information hidden in the data by means of data mining technologies. In this paper, we make use of some data mining technologies and methods meeting the characteristics of astronomical data to automatically classify celestial objects and estimate photometric redshifts. The main contributions are following:

(1) Estimating photometric redshifts with artificial neural networks

Photometric redshifts have their own developed predominance, especially in terms of the large sample of faint objects difficult to obtain their spectra, have much more statistical meaning, which enables us to gain a deeper understanding of cosmic large-scale structure and the formation and evolution of galaxies. In this paper, we apply the artificial neural network of data mining technologies to estimate photometric redshifts for the Sloan Digital Sky Survey (SDSS) data. Three main effective magnitude parameters: namely Petrosian magnitude, model magnitude and dereddening magnitude, along with galaxy color indexes and the Petrosian 50 and 90 percent flux radii in five bands, have been applied in our experiments. By the contrast, we obtained the best set of parameters for neural networks used for estimating photometric redshifts, which provides the reference of feature selection for astronomers.

(2) *K*-Nearest Neighbors for automated classification of celestial objects

K-Nearest Neighbors (*k*NN) algorithm is one of the simplest and most flexible and effective classification algorithms in data mining technologies, which has been widely used in many fields. Here we use *k*NN method to realize automatic classification of astronomical objects. Based on the difference of physical properties of different types of objects in different wavelengths, we try to separate AGNs from stars and galaxies in the multidimensional parameter space. The samples are obtained from different wavelength catalogs by positional cross-identification to extract various parameters which contain different information for these sources. The experimental result shows that the accuracy

of classification adds up to 97.73%, which suggests that automated classification using *k*NN method is reasonable and applicable.

(3) The separation of morphology types and the estimation of photometric redshifts for galaxies

We put forward a new scheme to study how to improve the accuracy of photometric redshifts. At first, the total galaxy sample is separated into two kinds, namely early-type and late-type galaxies, by means of *k*-means automated clustering algorithm. Then photometric redshifts for both kinds of subclasses are estimated with artificial neural networks, respectively. Finally we calculated their combinative accuracy and compared the predicted accuracy with that using the total sample alone. The comparison shows the former gave better results than the latter and the best rms dispersion amounts to 0.0192. Especially, the rms scatter of early-type galaxies approaches 0.0164. Our experiment indicates that the proposed scheme is reasonable, feasible and effective.

Keywords: data mining, neural networks, automatic classification, photometric redshifts

目　录

插 图 目 录

表 格 目 录

第一章　绪论

1.1　引言

人类在漫长历史长河中，永远依存着自然环境。回顾人类的自然科学发展史，宇宙的影响无时无刻不在进行着。换句话说，天文学在各种自然科学的发展过程中，始终起着先导性的作用。天文学的研究成果不仅能直接服务于电波通讯、宇航、军事、航海、大地测量、地球环境变化、灾害研究等的发展和国民经济与国防建设，而且不断地加深着人类对自身所处的宇宙环境的认识，成为人类文明进步的重要标志。

众所周知，天文学是研究地球以外的天体及其毗邻环境的科学，它的研究对象包括太阳系天体、恒星、银河系、河外星系、行星际和星际物质，直至整个宇宙。它以各类现代尖端技术作为观测手段，收集和处理来自宇宙的全波段电磁辐射和其他信息，应用物理学、数学、化学等基础学科的理论和方法，研究各类天体、天体系统乃至整个可观测宇宙的运动规律、化学组成、物理性质和演化过程，以及人类生存的天体-地球及其空间环境。天文学是当今科学最具活力的一个分支，是基础创新力的源泉，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。

天文学是依赖于观测的基础研究科学，观测设备的先进程度就决定了天文学的发展水平。因而各个发达国家都在投入巨额资金竞相独立或合作研制新一代地基和空间大口径望远镜。如美国口径为 10 米的 Keck I 和 Keck II 以及相应的光学干涉仪、欧洲的 16 米 VLT 及相应的干涉仪、日本的 8.2 米 SUBARU 等。高光效大面积 CCD 以及大视场多目标光谱仪的出现，又使得天文学在深度和细度上正朝着前所未有的方向发展。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，天文学经历了巨大的飞跃，也迅速推进着人类对宇宙的认识。

近年来，国际上涌现出了大批新一代望远镜巡天观测项目，在它们的推动下，使得天文数据量在急剧地增长。SDSS巡天 (the Sloan Digital Sky Survey，简称SDSS)是目前巡天中最富野心的巡天计划，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，为星系、类星体和恒星的大型光谱巡天奠定了基础。2MASS巡天(the Two Micron All-Sky Survey，简称2MASS)是一个近红外（J、H和K_s波段）的全天巡天项目。FIRST巡天(the Faint Images of the Radio Sky at Twenty-cm Survey，简称FIRST)是一个 1993 年开始运行的射电波段巡天，覆盖南北银极近一万平方度。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，简称LAMOST)于 1997 年正式立项，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。

各大巡天项目的投入运行及各种新技术的应用，使得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。天文观测也从可见光、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，从而使天文学发展到了一个全新的阶段，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。来自各个波段的数据量以指数量级增长，以 TB 量级甚至 PB 量级计量。但是我们不可因为获得如此巨大的数据量而沾沾自喜，如果不能有效地利用、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好几个 TB 的数据？如何从具有十几亿甚至几百亿的天体或天文数据中进行科学发现？这是摆在天文学家面前不可回避的问题。为了有效地解决这些问题，应对形势发展的需要，引入适合天文发展和需要的数据挖掘与知识发现技术是必需的也是必要的，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或天文现象，这将有助于推动天文学理论的进一步发展和完善。

1.2 数据挖掘技术

数据挖掘（Data Mining）技术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、统计上重要的结构和事件。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。随着新的巡天带来了巨量的科学数据，对这些巡天数据的联合使用和分析，挖掘出有用的信息，将涌现出无法预见的、意义重大的科学发现。

数据挖掘可粗略地分为三步[1]：数据准备（data preparation）、数据挖掘，以及结果的解释与评估。其过程如图1-1所示：

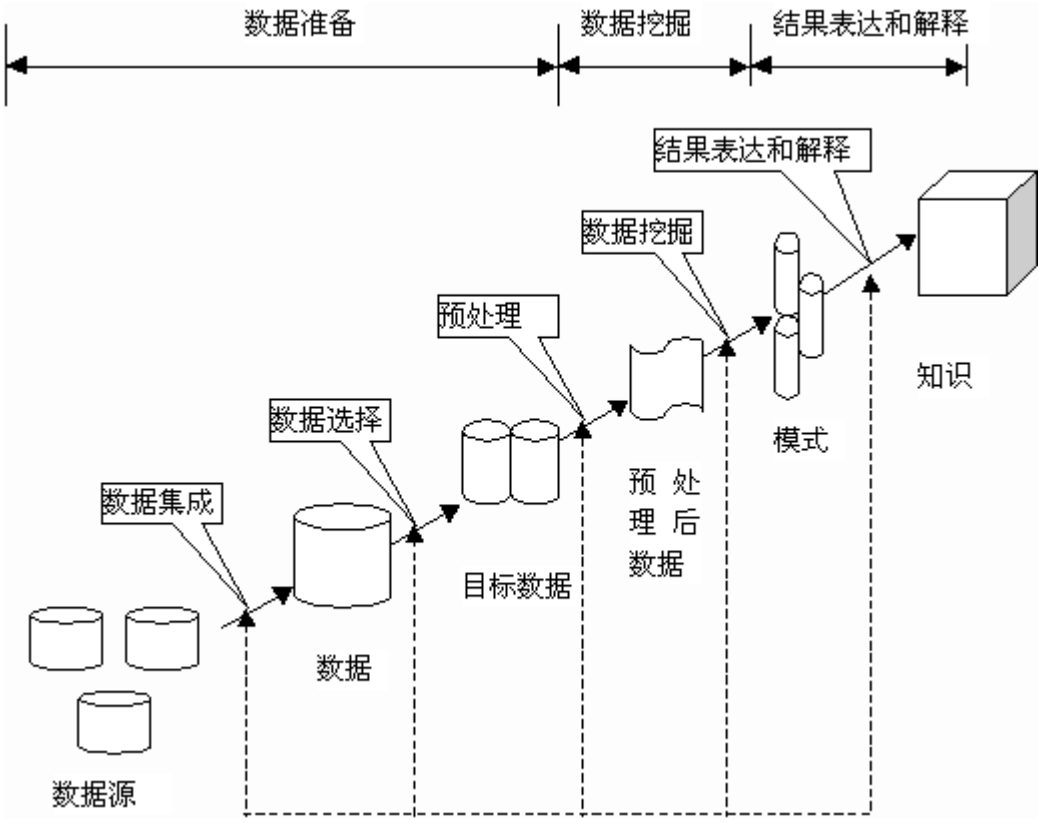

图 1-1　数据挖掘的流程图

（1） 数据准备

数据准备就是对被挖掘的数据进行定义、处理和表示，以使它适应于特定的数据挖掘方法。数据准备是数据挖掘过程中的第一个重要步骤，在整个数据挖掘过程中起着举足轻重的作用。它主要包括如下三个过程：

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。提高数据的质量，为进一步的分析做准备，并确定将要进行的挖掘操作的类型。

③ 数据的转换

找到数据的特征表示，用维变换或转换方法，减少变量的数目或找到数据的不变式。将数据转换成一个针对挖掘算法建立的分析模型，而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。

（2） 数据挖掘

对所得到的经过转换的数据进行挖掘。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。除了完善所选择合适的挖掘算法外，其余一切工作将会自动地完成。然后搜索并产生一个特定的感兴趣的模式或一个特定的数据集。

（3） 结果的解释与评估

解释并评估结果。解释某个发现的模式，去掉多余的无关的模式，转化为某个有用的模式，以使用户能够理解。其使用的分析方法一般由数据挖掘操作决定，通常会用到可视化技术。

数据挖掘的过程与人类问题求解的过程存在巨大相似性。挖掘过程可能需要多次的循环反复，每一个步骤一旦与预期目标不符，都要回到前面的步骤，重新调整，重新执行，因此好的数据是数据挖掘的关键。数据挖掘是多技术的综合，如数据管理、机器学习、统计推理、高性能计算、决策支持、可视化等。

1.3 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、机器学习等。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。

数据挖掘的任务是从数据中发现预测型(Predictive)和描述型(Descriptive)模式。而按实际作用可分为以下 6 种：

（1）分类模式(Classification)：分类模式是把数据集中的数据项映射到某个给定的类上。首先从数据中选出已经分好类的训练集，在该训练集上运用数据挖掘的分类技术，建立分类模型，对于没有分类的数据作出预测分类。注意这里类的个数是确定的，预先定义好的。即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，然后用这个分类规则对数据库中的其它记录进行分类。常用算法有 *K* 近邻算法、决策树法、神经网络法、径向基函数法等。

（2）回归模式（Regression）：回归模式的函数定义与分类模式相似，其差别在于分类模式的预测值是离散的，而回归模式的预测值是连续的。

（3）关联模式(Association)：关联模式是数据项之间的关联规则。关联规则具有如下形式的一种规则：A=>B。利用规则归纳方法进行数据挖掘，其目的是挖掘隐藏在数据间的相互关系。目前有许多较成熟的算法，如：APRIORI、DHP 等。

（4）聚类模式(Clustering)：聚类是一种对具有共同趋势和模式的数据元组进行分组的方法。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，确定每个记录所在的组别。分组后，组与组之间被认为是相异的，而组内记录被认为具有相似性，针对不同的组可制定不同的策略。聚集是对记录分组，把相似的记录在一个聚集里。聚类和分类的区别是聚类不依赖于预先定义好的类，即不需要训练集。聚类方法有 *k-means* 算法、分层凝聚法、动态聚类法、模糊聚类法等。

（5）时序模式(Sequential Patterns)：时序模式根据数据随时间变化的趋势，发现某一时间段内数据的相关处理模型，预测将来可能出现值的分布。时序模式可看成是一种特定的关联模型，它在关联模型中增加了时间属性。时序模式注重于从时间上分析数据间的前后序列关系，找出重复发生概率较高的模式。

（6）孤立点模式(Outlier Analysis)：用距离的观点分析那些远离高密度区、大部分点域的点，以发现异常的行为。通过数据挖掘发现的知识可以用于信息管理、查询优化、决策支持、过程控制、数据自身的维护等。

在数据挖掘任务中，分类模式和回归模式是使用最普遍的模式，主要用于预测。分类模式、回归模式、时序模式也被认为是受监督的，因为在建立模式前数据的结果是已知的，可以直接用来检测模式的准确性，模式的产生是在受监督的情况下进行的，一般在建立这些模式时，使用一部分数据作为训练样本，用另一部分数据来检验、校正模式。聚类模式、关联模式则是非监督的，结果在模式建立前是未知的，模式的产生不受任何监督。

数据挖掘的方法可分为：机器学习方法、统计方法、神经网络方法和数据库方法。机器学习又可细分为：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。神经网络方法可细分为：前向神经网络（*BP* 算法等）、径向基神经网络、自组织神经网络（自组织特征映射、竞争学习等）等。一般来说，数据挖掘利用的技术方法越多，得出的结果精确性就越高。原因很简单，对于某一种技术不适用的问题，使用其它方法则有可能奏效。

数据挖掘技术可以有效地应用于天文学中，有效地处理天文学面临的“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，从而促进天文学的发展。对于数据挖掘的应用，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、回归分析或聚类的问题。近年来，这方面的研究已成为天文数据研究领域的热点。本文将对这几方面进行研究，分别使用数据挖掘中的分类、回归和聚类

模式在天文数据中挖掘我们感兴趣的信息，使用的数据挖掘方法包括 *K* 近邻算法、多层神经网络和 *k-means* 算法，侧重于智能化工具—神经网络方法的应用。

第二章　神经网络

从十多年前，天文学开始发生革命性的变化，这一变化是由前所未有的技术进步所推动的，即望远镜的设计和制造、大尺寸探测器阵列的开发、计算能力的指数增长以及互联网络的飞速发展。在国际天文界的共同努力下，各个巡天项目所观测的海量天文数据将能很快地通过互联网发布出去，成为全球共用的资源。如何有效而科学地加工、处理和分析这些数据，已成为摆在天文学家面前不可回避的问题。要对这些数据进行科学有效地处理，就需要一整套能对大量的数据进行从预处理、特征提取/选择、数据挖掘到对结果的解释与评估等一系列工作的全自动化的工具。由于天文数据本身具有的复杂性（如非线性、持续性、高维性以及普遍存在噪声和缺值等），传统的数据分析方法已显得力不从心，因此需要一些新的算法和工具来解决这些难题。于是以神经网络、模糊近似、遗传算法等为基础的人工智能工具应运而生。神经网络(Neural Network，NN)，又称人工神经网络(Artificial Neural Network，ANN)，它具有很强的适用于复杂环境和多目标控制要求的能力和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领域。目前神经网络在模式识别、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了广泛的应用。同样它在天文领域也有着广泛而成功的应用。

2.1 神经网络原理

神经网络最初是作为人脑功能的简单模型而引入的（其中节点代替神经元，多层连接代替树突和轴突），尽管神经网络包含的神经元数目远不及人脑，但对信号的逻辑处理过程与人脑非常相似。它是一种在对人脑组织结构和运行机制的认识理解基础之上，模拟大脑结构和智能行为的信息处理系统，是由大量神经元广泛互连而成的复杂网络，这些神经元用于处理神经网络中传递的信息，并通过权值连接起来。一个神经元可以接受与它相连接

的所有神经元输出的信息，作为它的输入，使用激活函数计算出相应的输出，再将输出传递给其它神经元。具体过程如下：对于一个 N 层神经网络，其第一层（$m=1$）是输入层，中间一层或多层为隐含层，最后一层($m=N$)为输出层。神经网络结构如图 2-1 所示。

输入层的n个神经元x_i组成n维输入向量x，x_i分别以权值$w_{j,i}^{(1)}$连接到第一隐含层，隐含层神经元得到输入向量x的加权和，通过线性或非线性激励函数在第j个神经元处产生的输出值：

$$z_j^{(1)} = f(\sum_{i=0}^{n} w_{j,i}^{(1)} x_i) \qquad (2\text{-}1)$$

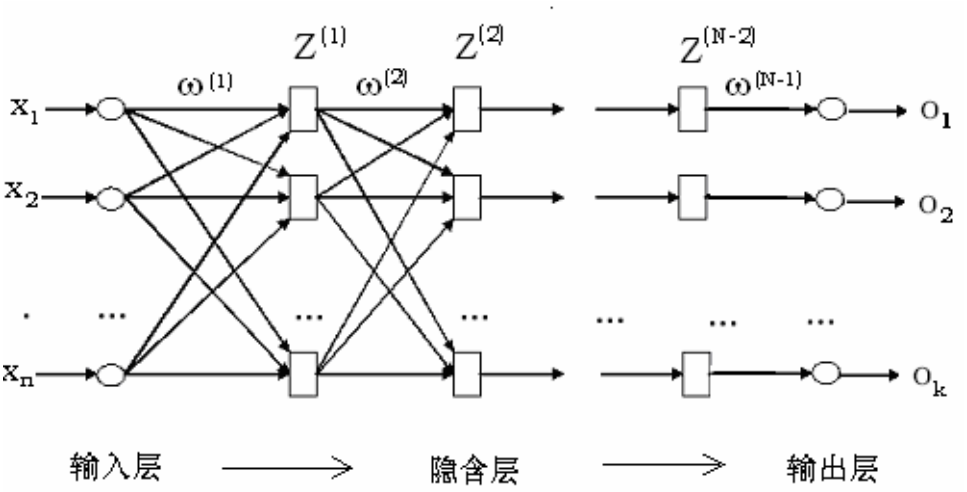

图 2-1　神经网络结构图

上角括号中数字分别表示隐含层和连接权值的层次，$w_{j,i}^{(1)}$是到第一隐含层（从输入层的第i 个神经元x_i到第一隐含层中第j个神经元）的连接权值（当$i=0$ 时，附加常量$w_{j,0}^{(1)}$表示第j个神经元的偏置，x_0取为 1，除了有固定不变的输入值 1 以外，其他类似权重）。依次类推，$w_{j,i}^{(k)}$则是到第m隐含层（从第（$k-1$）隐含层的第i 个神经元到第k隐含层中第j个神经元）的连接权值，f是激励函数(如$sigmoidal$函数、$tansig$函数、$purelin$函数等)。输出值$z_j^{(1)}$又以权值连接到第 2 隐含层，作为该层神经元的输入，如

此循环直至最后的隐含层，而输出层的各神经元得到第$N-2$隐含层各神经元的输出值$z_j^{(N-2)}$加权输入求和，通过激励函数在第k个神经元产生的输出值为：

$$o_k = g(\sum_{j=0}^{M} w_{k,j}^{(N-1)} z_j^{(N-2)}) \tag{2-2}$$

M是最后（第$N-2$个）隐含层神经元的个数，$w_{k,j}^{(N-1)}$是从最后隐含层第j个神经元到输出层第k个神经元的连接权值（即加权系数），g是输出层的激励函数。神经网络的自由参数是权向量，其所含层数、每层神经元数及激励函数都是最初选择的，它们一旦选定也就确定了神经网络的结构。加权系数反映了神经系统中神经元的突触强度，它可以加强或减弱上一个神经元的输出对下一个神经元的刺激。实际上，神经网络的工作就是要首先给定一个有代表性的样本数据，通过“训练（training）”过程不断地调节各层的连接权值以最大程度地减小输出误差，优化网络。从而使网络能够反映出样本的内在规律并具有一定的泛化能力，对于一个给定的输入可以产生合适的输出值[2]。

人工神经网络的特性取决于所采用的学习算法和网络模型。在神经网络中，修改权值的规则称为学习算法（learning algorithm），可以定性地分成以下三类：

（1）监督方法（Supervised methods）：也称有教师学习，在训练过程中预先给出目标输出，根据实际输出与期望输出的误差来调整网络权值。监督方法通常速度快而准确，但是要构造一个合适的训练网可能是相当麻烦的。

（2）非监督方法（Unsupervised methods）：又称无教师学习，系统完全按照外部提供数据的某些统计规律来调节自身的参数或结构，以反映外部输入信息的某些固有特性。这种方法常用来对数据进行压缩（如降维或聚类）。

（3）再激励式学习：是指外部环境对系统输出结果只给出评价信息（奖励或惩罚），而不是给出正确答案，系统通过强化那些受奖励的结果而学习[3]。

神经网络的拓扑结构主要有以下两种：

（1） 前馈网络：即信息仅从 K 层向 $K+1$ 层传播，并不循环，它能对输入作出较快速的反应。

（2）反馈（或循环）网络：信息可以循环传播，每次只处理一个输入，网络要循环多次才能作出反应。

具体选择哪种神经网络的结构和操作模型完全依待解决问题的内部特性而定，神经网络通过不断地尝试和纠错过程，从而产生一个好的处理方法。为了使神经网络技术能有效地工作，所有的网络需要经过长期的优化过程，对输入数据的噪声和不准确性强度也要进行测试。

2.2 神经网络适用于天文学应用的主要特征

神经网络之所以能在天文数据处理中得到广泛应用，在于神经网络具有能最大限度地反映天文数据本身的复杂性的许多特点：

（1） 逼近非线性映射关系的能力：已有理论证明任意的连续非线性函数映射关系都可由某个多层神经网络以任意精度加以逼近。这种组成单元简单、结构有序的模型是非线性系统建模的有效框架，预示着神经网络在处理属性之间存在非线性关系的天文数据时具有很好的应用前景。

（2） 对信息的并行分布存储方式：神经网络的大规模互连网络结构及对信息的并行分布式存储，克服了传统的串联工作方式，使其在对海量的天文数据处理时具有一定的优势。

（3） 高强的容错能力：神经网络的并行处理机制及冗余结构特性使其具有较强的容错特性，提高了信息处理的可靠性和鲁棒性，能够对天文数据不可避免的缺值现象应付自如。

（4） 对学习结果的泛化和自适应能力：经过适当训练的神经网络具有潜在的自适应模式匹配功能，能对所学信息加以分布式存储和泛化，这是其智能特性的重要体现，也是使其成为处理巨大天文数据库的一种智能化工具的原因之一。

2.3 神经网络在天文中的主要应用模型

神经网络在天文应用中主要有以下几种模型：

（1） 前向神经网络：是一种分层排列的网络结构，每一层的神经元输出只和下一层神经元相连。这种网络结构特别适用于BP 算法(Back Propagation，BP)，它以监督的方式进行学习。对于网络，只要给出有代表性的训练子网，经过训练后，网络就具有很强的泛化能力，使得对任意未知数据都可以做出预测。它具有执行快和对非线性逼近较容易等优点，目前在天文中已得到了非常广泛的应用，主要用于测光红移[4]、星系形态分类[5]等方面。由于前向网络结构具有很强的互连能力，也常应用在对天文望远镜的系统控制中[6]。但其缺点是学习收敛速度太慢、网络的学习记忆具有不稳定性和对参数的选择较为敏感。

（2） 自组织映射网络(Self-Organizing Maps，SOM)：在这种网络结构中，同一层之间存在着相互关联，神经元之间有相互制约的关系，但从层与层之间的关系来看还是前馈式的网络结构。它采用非监督竞争的学习方式，训练完全由数据驱动，各神经元竞争对输入模式的响应机会，最后仅一个神经元可成为竞争的获胜者，并将那些有关的连接权值朝着更有利于获胜神经元竞争的方向调整，这一获胜神经元就反映了对输入模式的分类[3]。对于一个给定的高维数据，自组织映射网络可以利用神经网络的非线性关系把它投影到一维或二维空间上，并尽可能保持原始数据空间的拓扑结构，实现对数据的降维，使之更具有可视性，以便能更好地理解高维数据。自组织映射网络在天文学中主要用于分类，如恒星与星系的分类[7]、图像分类[8]、光谱分类[9]、伽玛射线暴分类[10]、星表分类[11]等。但 SOM 也存在一定的缺陷：因为它仅以输出层的单个神经元来代表某一类模式。所以一旦输出层中某个输出神经元损坏，则会导致该神经元所代表的该模式信息全部丢失。同时由于对输入数据维数的减少，SOM 扭曲了输入数据空间的描述并掩盖了数据样本间的相互关系。

（3） 径向基函数神经网络（Radial Basis Function，RBF）：网络由三层组成，输入节点只传递输入信号到隐含层，隐含层节点由与高斯核函数类似的辐射状作用函数构成，而输出层节点通

常是简单的线性函数。隐含层节点中的作用函数（基函数）对输入信号将局部产生响应，也就是说，当输入信号靠近基函数的中央范围时，隐含层节点将产生较大的输出，由此看出这种网络具有局部逼近能力。其连接权的学习修正仍可采用 BP 算法，从理论上而言，RBF 网络和 BP 网络（使用 BP 算法的前向神经网络简称 BP 网络）一样可近似任何连续非线性函数，两者的主要区别在于使用不同的作用函数，BP 网络中的隐层节点使用的是 *Sigmoid* 函数，其函数值在输入空间无限大的范围内为非零值，而 RBF 网络的作用函数则是局部的[2]。RBF 网络采用保证全局收敛的线性优化算法，并且还具有唯一最佳逼近点的优点，因此 RBF 网络以其结构简单，训练过程快速等优点在许多领域取得了巨大的成功。在天文学中应用于光谱的自动分类、光谱识别率的提高[12]、恒星和星系的自动分类[13]等方面。

（4） Hopfield 神经网络：是一种反馈回归网络，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。网络中的神经元在 *t* 时刻的输出状态实际上间接地与自己 *t-1* 时刻的输出状态有关。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时，它的能量函数最小。这里的能量函数不是物理意义上的能量函数，只是在表达形式上与物理意义上的能量函数概念一致、可表征网络状态的变化趋势、可以依据 Hopfield 工作运行规则不断进行状态变化并最终能够达到某个极小值的目标函数[14]。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。其在天文学中应用于图像恢复[15]等工作。

（5） PCA 神经网络（Principal Component Analysis Neural Network，PCA NN）：一般仅有两层（输入层和输出层）的前馈神经网络，每一个输入加权后代入激活函数计算输出，再按照一定的规则修改权值，不断地学习，直至满足网络的要求才结束。根据学习阶段神经元间的反馈连接形式，可以把网络结构分为分层式和系统式。分层式修改权值时逐级求和；而系统式则为整体

求和。对于非线性问题可以用非线性 PCA 神经网络或强 PCA 神经网络(robust PCA Neural Network，robust PCA NN)来处理。这些方法直接对数据处理，从中提取特征矢量，大大减少了运算量，因此在对高维天文数据进行预处理时，可以很好地实现对数据降维、特征向量提取、图像压缩和目标探测[16]等。

2.4 神经网络在天文中的应用

由于神经网络非常适用于处理数据的非线性复杂关系，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，以不同的网络模型分别实现数据的聚类、分类、关联、回归、模式识别等多种任务。关于几种主要的神经网络模型的优缺点及其在天文学中的应用如表 1-1 所示。

有关神经网络在天文学中应用的详细介绍可参见文献[17~19]。下面将详细介绍一些神经网络在天文中的具体应用事例。

表 1-1 几种网络模型的优缺点及其在天文中的应用

神经网络模型	优 点	缺 点	天文应用
前向神经网络 (MLP 和 BP 算法)	主要采用监督方法。通过并行的方式解决非线性问题；通过误差反传以监督的方式训练权值；泛化能力强。	采用梯度下降法，学习速度慢；对复杂非线性问题训练出来的参数容易陷于局部极小；对参数选择敏感；隐层节点的选取无理论指导。	测光红移[4] 星系形态分类[5] 系统控制[6]
自组竞争神经网络（SOM）	具有无监督性、自动提取特征以及主分量分析、聚类、编码和特征映射等功能，有助于可视化，能够实现数据降维，适用于数据量化。	执行较慢；用于降维时会扭曲数据间的关系；当输入模式较少时，分类结果依赖于模式输入的先后次序；网络在没有经过完整的重新学习之前，不能加入新的类别。	恒星/星系分类[7] 图像分割[8] 光谱分类[9] 星表分类[10]
径向基函数神经网络（RBF）	采用监督方法，网络结构简单，收敛速度快，避免局部极小；非线性局部逼近/全局逼近能力和拟合能力强。	基函数中心难以选择；对大样本处理能力较弱，如：需要复杂的结构和训练算法且收敛速度慢。	光谱自动分类[12] 恒星/星系分类[13]
Hopfield 神经网络	监督学习，可用于联想记忆，处理优化问题等。	能量函数一旦陷于局部极小值，就不能自动跳出，从而不能达到全局最小值，这样无法求得网络最优解。	天文图像恢复[15]
PCA 神经网络	具有自组织、自学习特性，有助于提取特征矢量，可以用于降维、除噪，从而减少运算量。	对弱信号分辨能力弱，而且不适合直接处理非线性问题；对于非线性问题最好用非线性 PCA 神经网络或强 PCA 神经网络。	特征提取[16] 目标探测[16] 恒星光变曲线分析[20][2]

2.4.1 图像识别

图像识别是把图像分成若干个各具特性的区域并提取感兴趣目标的技术过程。天文图像的每一个像素值都是“背景”信号和天体的光信号之叠加，因此需要首先进行图像识别。标准的图像识别方法不适宜处理具有特殊性质的天文图像，因为天文图像有的具有上千个亮度等级、多个亮源、不同尺度的微弱弥散天体和平滑背景等复杂情况，同时极暗天体的图像、探测器电子性质及仪器的高斯白噪声等更增加了图像识别的复杂性。天文学家通常直接利用图像上出现的物体或从诊断图中衍生的参数对天体分类，然而这种方法耗时长，需要观测者事先知道“如何做”，观测者的经验也会影响分类结果，从而存在着很多问题，如：合适参数的选择因人而异，它们之间不能或者很难比较；将三维或更高维投影到二维图上，不管问题多复杂，通常也只能考虑二维特性。随着人工智能研究的发展，出现了一些新的技术和方法来解决这些问题，神经网络就属于其中的一种。

为了解决天文图像识别问题，科学家们进行了广泛的研究。Bertin 和 Arnouts 提出了一种自动地从天文图像中提取源的技术 SExtractor[21]，利用训练多层前向神经网络来模拟天文图像，能够很好地完成对恒星和星系图像的识别。SExtractor 的显著特性是能对大图像中各种形状和星等的天体进行自动处理，因此特别适合于分析大的河外星系观测图像。Andreon 等人发展了神经网络软件包 NExt，能够自动地对天体探测、对恒星/星系分类。NExt 先用非线性主成分分析提取特征，将每个像素投影到主矢量空间中，把获得的特征向量值作为非监督神经网络的输入值，通过非监督神经网络把像素分为物体和背景两类，再用一套参数来识别重叠物体并把它们分解，最终使物体轮廓规则化，使图像更加清晰[16]。

Jorge 把神经网络与小波变换结合对天体图像进行识别[22]。因为天文图像有不同的亮度等级，Jorge 先用小波变换对图像进行多分辨率分析，分解出恒星和明显天体，去除了噪声，再用自组织映射网络（SOM）对其余部分（包括展源和背景）进行分解。这种方法对含有噪声的数据表现出明显的灵活性，避免了结果的

过度散碎。神经网络这种模式识别技术在天文学大型的数字巡天观测的图像识别和目标探测中已经越来越显示出其优越性，取得了较好的效果。

2.4.2 恒星/星系分类

神经网络在天文数据挖掘应用中最常见的就是恒星/星系的分类。天文学家已经使用神经网络在恒星/星系分类中做了大量的工作，取得了一定的成功。Andreon 等人利用多层感知机神经网络（Multi-layer Perceptron, MLP）对恒星和星系分类，他们将数据分成三个独立的子集，分别用于训练、评价和测试。先对数据进行特征提取，再用提取的特征训练 MLP 神经网络，通过不断训练，从而使学习达到最优化，最后来评价训练结果，以保证其适用于整个数据[16]。Miller 利用自组织映射（SOM）对恒星和星系进行分类，将得到的结果与人工分类获得的结果进行比较，得出自组织神经网络对于极限星等为 20 星等的星系分类准确率为 98%，达到了其他分类方法的准确率，但在训练时不需要大量的人工参与，实现了分类的自动化[7]。

2.4.3 光谱分类

光谱分类是恒星和星系天文学中的一个重要的研究课题。通常的恒星光谱分类是由专家采用人工或半人工的方法将未知谱型的恒星光谱与标准的恒星光谱作比较，从而获得恒星的光谱型。这种方法效率低，受个人主观因素影响也比较大，难以形成统一的标准。随着巡天项目海量数据的出现，迫切需要高效且准确的光谱自动处理系统。神经网络作为一种重要的人工智能工具，已经在这方面得到了广泛应用。

Vieira 分别用监督 BP 网络和非监督的自组织映射（SOM）对 IUE 标准星的低色散光谱进行分类[23]，并与人工分类做了比较，结果发现差错率很小，表明这两种方法都能实现真正意义上的光谱数据自动分类，而不需要人工分析光谱特性，因此对于处理海量数据具有很大的意义。Bailer-Jones 研究了利用神经网络

对 MK 光谱进行自动分类，对于包含多余信息的恒星光谱数据，先用主成分分析（Principal Component Analysis，PCA）方法对其进行压缩，消除噪声，分离出伪光谱，再将处理后的数据作为神经网络的输入，使用多层监督的神经网络结构将 MK 的光谱按光度等级进行分类，对矮星和巨星的分类准确率达到了 95%，说明神经网络完全可以胜任对整个光学波段的光谱进行较精确的分类[24]。

姜玉刚等人结合了小波变换与径向基函数神经网络对光谱自动分类[12]，其识别率比PCA特征提取结合RBF神经网络的分类方法提高了很多，而且其算法具有很好的鲁棒性，降噪能力也很强，非常适合信噪比较低的天体光谱的自动识别，具有很高的应用价值。Bai等人把Kalman过滤器与径向基函数神经网络结合对恒星光谱进行自动分类[13]。此前，Snider等人曾用神经网络方法对贫金属星的三维光谱进行过分类[25]。随着多种方法与神经网络方法的结合使用，光谱自动分类问题将得到有效的解决，分类的速度和准确率也大有提高。

2.4.4 星系的形态分类

星系的形态分类在宇宙大尺度研究中很重要，它有助于更好地给出星系模型，使我们更易于理解星系的结构及其演化过程。对星系形态分类通常是观察照相底片，然而这决不是件容易的工作，要求有一定的技能和专业知识。目前随着哈勃深场望远镜和大规模巡天项目获得的河外星系数据越来越多，如何从这些数据库中对星系进行识别和分类是当前天文学家面临的严峻挑战，幸运的是计算机的发展和人工智能的出现使得问题能够得以解决。

Goderya 提出了三种建立星系自动分类系统的方案，其中一种是以前从未采用过的新方案，该方案使用计算机提取星系外型特征，再利用人工神经网络建立自动分类系统对星系形态进行分类。同时他给出了此种方案的一个原型[26]。Naim 等人基于星系蓝波段盘区表现出的形态对星系进行分类，他们用神经网络自动地提取图像数据，获取星系的膨胀尺度和悬臂数等形态特征，并用这些特征来训练和测试网络，从而实现对星系形态的分类。他

将获得的结果与六位专家的分类结果相对比，发现差错率非常小[27]，因此神经网络可以很好地应用于星系形态的分类。

2.4.5 测光红移评估

星系红移（即星系离开观测者的退行速度）在观测宇宙学理论中极为重要。红移既可用光谱数据也可用测光数据来测量，用光谱数据虽然测得的红移准确但耗时长；测光数据虽不够准确并有系统误差，但较容易应用于大的样本。因此目前对于大样本，可使用来自同一仪器或系统的多波段的测光数据，并可用光谱红移作为子网，使用监督神经网络对测光红移进行评估。

Vanzella等人使用多层感知机神经网络（MLP）来测定红移。以颜色和面亮度作为参数，应用每层有不同数量神经元的三层或四层网络结构，预测多色星表HDF-S的测光红移，将其结果与已有文献中的光谱红移值作对比，他们得出在 $0.1<z<3.5$ 的红移范围内测光红移和光谱红移吻合得较好。他们又对SDSS第一次释放的大样本光谱数据在小红移范围（$0<z<0.4$）内进行预测，同样得出较好的结果。此方法应用于大样本数据库时计算速度很快，如对 10^5 个星系进行红移估测，用Pentium 3、1.1GHz的计算机仅用了几秒[4]。

Collister提出了一种使用人工神经网络来评估测光红移的软件包ANNz[28]，它能从红移已知的训练网中学习红移和测光数据的关系，在存在大量有代表性的训练网的情况下，ANNz比传统的模版匹配方法更好，更占优势。ANNz软件包在处理SDSS 第一次释放的数据的过程中得到肯定，当红移范围约为 $0\leq z\leq 0.7$ 时，确定的红移值剩余标准偏差为 0.023。

2.4.6 时间序列分析

寻找在时间上具有周期性、空间上独立的信号是许多研究领域的重要研究内容。在天文学中，随着观测仪器灵敏度的提高，越来越频繁地发现在某时间段内以前被认为是常量的信号实质

是变化的，这些信号分别是在某一特定时间段内提取的，属于非均匀信号，而这些信号对于研究如变星、活动星系核等之类的天体特别重要。我们需要找出这些变化信号（如非均匀样本光变曲线）的周期。典型的谱线分析方法对于处理非均匀样本无能为力，而神经网络技术能够很好地解决这些问题并取得了成功。

Tagliaferri 提出了以神经网络为基础的多信号频率探测器 MUSIC[20]，使用非监督的 Hebbian 非线性学习算法提取非均匀样本信号的频率，由于神经网络对数据的处理方式是并行、分布式的，因此对含噪声的信号具有很强的容忍性。对非均匀样本信号进行光谱分析可分为三个阶段：对输入数据归一化进行信号预处理；用非线性 PCA 神经网络提取主要特征向量；应用多信号分类算法来提取信号频率。相对于周期图来说，这种方法用于处理非均匀信号具有很大的优势：简单直接，对频率间隔不敏感。

2.4.7 其他应用

目前神经网络技术在天文领域中的应用已经相当广泛，除了上面所介绍的应用外，还有许多重要应用，如对太阳耀斑进行自动探测[29]、模拟太阳风和内磁场的统计分布[30]、自动确定恒星大气参数[31] 、对类星体（QSO）、极亮红外星系（ultraluminous IR galaxies）和伽玛射线暴等特殊天体的识别[32]、多频率模拟数据的分解[33]、神经网络对干涉仪输出的已知源信号噪声白化[34]、多波段数据的分类[35]、行星际活动探测[36]、从宇宙线照射引起的强子中分辨出伽玛射线[37]、CCD 缺陷的探测和分类[38]、来自物端棱镜图像光谱的探测和提取[39]、视宁静度和温度的预报[40]、望远镜导星[41]、波前参数探测[42]、望远镜进度表[43]、行星状星云的分类[44]等。

2.5 总结与展望

综上所述，由于神经网络对信息进行分布式存储和并行处理，具有很强的非线性映射能力、良好的适应性和容错性，与一

些传统算法相比，有其独特的优点，在天文学中得到了广泛的应用。针对目前天文学“数据雪崩”及天文数据的非线性、噪声普遍存在等现象，神经网络更能发挥出它的优势，越来越受到国内外天文学家的青睐，具有很好的发展和应用前景。

但神经网络犹如一个“黑箱”，我们只能看到它的输入和输出，其内部结构复杂，可解释性较差。即使是简单的神经网络也要由网络设计者给出网络的拓扑结构和激活函数，因此其不足之处是：使用的网络没有固定的结构（如网络的层数、激活函数的类型等），这些通常由研究者的经验来确定，并且网络的参数，也要由使用者在实践中不断地训练网络来确定，同时网络的训练时间很长等等，这些缺陷要求我们不断地对神经网络算法进行改进和优化，从而使网络不断地发展，使其越来越适用于解决具体问题。随着人工智能的出现和发展，人们可以充分利用多种算法，将其它算法与神经网络结合起来处理更加复杂的问题，从而避免了神经网络的缺陷。目前发展非常迅速的主要是以下几个方面：⑴ 将模糊算法和人工神经网络两者结合起来构建模糊神经网络系统；⑵ 基于遗传算法和模拟退火算法的人工神经网络系统；⑶ 将专家系统和人工神经网络结合起来的智能专家网络系统。这些前沿学科与神经网络有机的结合，必将给神经网络技术的发展注入新的活力，从而扩大人工神经网络在天文学中的应用，更好地处理天文领域的各种问题。

第三章 基于神经网络的测光红移

河外天体距离我们较远，它们的红移值较大；河内天体由其运动学特征决定了它们不可能有大的视向速度，其红移可以近似为零。我们知道天体的红移可以用多普勒效应来解释。1842 年奥地利人多普勒（Christian Johann Doppler）阐述了声学中的声调随声源运动而变化的原因，于是这种变化被称为“多普勒效应”。1848 年法国的斐佐（Armand Hippolyte Fizeau，1819-1896）利用多普勒效应来解释光源的运动，只要光源远离我们运动，光谱就会向红端移动（即“红移”），反之就会向紫端移动（“紫移”）；而且红移或紫移越大，光源移动的速度也越快。由于这种趋近和远离的运动都在观测者的视线方向上，故其运动速度称为“视向速度”。宇宙学中普遍存在的红移现象导致了大爆炸理论的产生，用多普勒效应来解释宇宙红移表明了现在可观测的宇宙是一个膨胀中的宇宙。1929 年，哈勃对河外星系的视向速度与距离的关系（当时只有 46 个河外星系的视向速度可以利用，而其中仅有 24 个推算出的距离。）进行了研究得出了著名的哈勃定律：

$$z = H \times \frac{r}{c} \qquad (3\text{-}1)$$

其中 z 是星系的红移，H 是哈勃常数，r 是星系到地球的距离，c 是光速。红移是所有河外天体的最重要的物理参数之一，通过红移可以计算出星系与我们的距离，这样我们就可以研究天体的各种物理性质，如：质量、大小、光度、爆发规模等。河外天体的红移也是天文学家研究星系的形成与演化和宇宙大尺度结构的基础。

3.1 测光红移

一般的星系红移是从星系光谱中分光证认出谱线（即强的发射线）来确定的，这就是所谓的光谱红移，它的精度相对较高。但对于多数遥远的星系，在达到狭缝或光纤光谱观测的极限时，很难得到它们的光谱数据，即使采用较长的曝光时间，所得到的光也会被分散，其信号的泊松噪声相对较大，因此采用光谱分光

测量来确定星系的红移成为一个难题。而目前普遍使用的电荷耦合器件（CCD）却可以捕捉到比分光观测极限暗得多的星系图像，这样的测光观测可以得到更深的极限星等，数据更可靠，因此使用测光数据估算红移值是获取大量遥远星系红移信息的有效手段。通过多色测光获得大量星系的红移样本，天文学家就可以用统计学方法从星系的数量和光度两方面来研究星系的演化。

1962 年，Baum 提出利用宽波带测光作为光谱能量分布（spectral energy distribution，SED）的近似来确定星系红移[45]。这种方法与光谱红移相比有很大的优势。首先，测光红移更高效、更能节省望远镜观测时间。对于光谱观测，要求源自星系的光在几个埃内被分割成窄带区域，这样每个窄带内仅获得星系总光度很小的一部分，因此要得到高信噪比信息就需要很长的整合时间；而对于测光来说，每个区域通常有上千埃，要达到相同的信噪比仅需要较短的曝光时间，并且宽波带滤光片所得到的信噪比要比分散的窄线光谱大很多。再者，图像探测器要比多目标光谱覆盖的天区更大，整个星系区域的图像可以同时拍到，而光谱观测却仅局限于狭缝或光纤定位的那些星系。这也意味着利用测光数据可以同时得到更多天体的红移值。使用测光红移方法对大星表估测出红移值，将会对宇宙大尺度结构元素（如：星系聚类、网状结构和宇宙模型）给予更好的约束。另外，根据大样本星系的红移可以对星系数密度演化、光度函数演化进行统计研究，将使我们更好地理解星系的形成与演化过程。

但是不足的是测光红移比光谱红移的不确定性更大，一般光谱红移的不确定度为 0.001，而测光红移则为 0.1。利用测光红移对单个星系研究可能是很不利的，但是要对大量星系的特征或宇宙大尺度结构研究来说，测光红移的不确定性又是可以容忍的，有时甚至会更加有效。在未来的若干年里大量暗弱天体仍将处于分光观测极限之外，因此测光红移是目前获取此类天体红移信息唯一可靠的途径，测光红移技术还被称为“穷人的红移机器”[46]。

从 Baum 提出测光红移后，很多方法已被应用于估计测光-红移关系。目前，它们被广泛地应用到了深视场和宽视场的多色巡天项目中，诸如 HDF、SDSS、 CADIS 等[47-49]。测光红移最

常用的方法是模板匹配，又称光谱能量分布（SED）拟合法，基于对光谱整体轮廓的拟合，即主要依赖于对 $Ly\alpha$、*Balmer* 跳变这类显著光谱特征的探测。这种方法是通过与从同一测光系统得到的光谱模板进行匹配获得的[50-51]。要求把每个星系的测光数据拟合为光谱能量分布（SED），构造包括所要研究星系类型、光度和红移等信息的模板光谱，通常是计算模板和实际光谱之间的 χ^2 值，使之最小化来确定红移。模板匹配技术充分利用了星系的 SED，实际上它可以很好地应用于很少或没有光谱红移的星系样本中。然而它的成功依赖于准确的具有代表性的 SED 模板的构造[52]。在预先有大量样本红移信息时，模板匹配技术不再是最好的方法。

测光红移的另一种方法是使用多项式或其他函数进行拟合，又称经验公式法。利用红移与星等（或颜色）间的经验关系，通过多项式拟合或其他函数拟合来确定样本的红移。实际上，它是将红移值拟合为测光数据的函数，既需要有测光数据又要有已知红移的大量代表性的样本。Connolly 和 Sowards-Emmerd 分别作了将红移拟合为星系颜色多项式的工作[53-54]，为了将测光红移和已知红移之间的拟合优化，多项式系数要不断地变化，这样对于未知光谱的星系，其测光红移就可以利用优化后的函数关系来确定。这种方法不需要知道星系光谱演化信息及进行诸多假设，使用起来比较简便。但它有一些缺点，譬如经验关系随不同观测系统的变化而变化，而且在高红移端，分光样本很少且不完备时，其红移估测将很不可靠。

近几年来，随着数据挖掘技术的出现，使得大样本星系的测光红移开始向自动化、智能化方向发展。一直以来，将各种数据挖掘技术应用于测光红移都是一个研究热点。天文学家开始考虑将新兴的智能化工具用于估计测光红移，包括支持向量机（Support Vector Machines，SVM）[55]，神经网络[56-58]等，其中神经网络是研究较多的一种，在实际应用中也取得了较好的效果。本文我们将利用神经网络从一个新的角度来估计测光红移，并探讨多种参数组合对红移测量精度的影响。

3.2 SDSS 巡天数据

3.2.1 SDSS 巡天介绍

本文使用的观测数据来自于 Sloan 数字巡天计划（Sloan Digital Sky Survey，简称 SDSS）。SDSS 是一项主要由美国 Alfred P. Sloan 基金会资助的、目前正在进行中的大面积的数字巡天计划。该计划开始设计于 1988 年，2004 年 4 月开始正式运行，预期共计 5 年。该巡天计划将预计覆盖北半天球的一半天区（北银极地区），以及少部分特选的南半天球天区[59]。该计划主要使用的望远镜为位于美国新墨西哥州的 Apache Point 天文台一个口径为 2.5 米的 3 度视场的专用望远镜。该望远镜配备有专用 CCD 照相机和光纤光谱仪，可分别进行大面积的测光和多天体的光谱观测。

SDSS 的 CCD 测光系统采用了对天体的漂移扫描技术，利用 6 组 CCD 同时对天体进行 5 个波段（*u,g,r,i,z*）的测光测量。所谓漂移扫描技术，是指望远镜的指向固定，而让天体由于地球的自转在成像 CCD 上漂移，最终通过扫描积分天体在 CCD 上的轨迹而得到天体的图像。天体漂移过单个 CCD 的时间，即每个天体的曝光时间为 54 秒。单块 CCD 的分辨率为 2048×2048 像素，每个像素的空间分辨率为 0.396 角秒。其 5 个波段（*u,g,r,i,z*）相应的中心波长分别为 22.0、22.2、22.2、21.3、20.5 等。最终，SDSS 的测光系统预计将能获得超过 1 亿个天体的准确的位置及星等测量。

SDSS的光谱系统采用光纤光谱技术，每个底片（*plate*）能同时跟踪 640 个候选目标。底片的视场半径为 1.49 度，光纤的孔径为 3 个角秒。任意两根光纤之间的最小间距为 55 角秒。光谱仪系统有两组，各负责 320 个目标，每组系统又分别由红端和蓝端两个光谱仪组成，红端和蓝端的分界在 6150 埃处。最终得到的每根光谱覆盖的波长范围为 3800 埃到 9200 埃，分辨率为 1800。每根光谱的平均曝光时间为 45 分钟，曝光所需要求是在 g=20.2 的情况下，每个像素上的信噪比满足 $(S/N)^2>15$。

SDSS 已经按计划公开释放数据，第一阶段观测为期五年。

累计观测的数据（SDSS-I）包括了 8000 个天区，测量了 675,000 星系、90,000 类星体和 185,000 星体的光谱数据。2001 年 5 月，SDSS 的早期数据（Early Data Release，简称 EDR）释放[60]；2003 年 5 月，SDSS 的第一批正式数据（Data Release One，简称 DR1）释放[61]。到 2006 年 6 月，SDSS 释放了第五批数据（DR5）。SDSS 的数据部分除了包括原始的图像和光谱之外。更重要的是提供了大量的观测目标的基本参数，比如测光中的星等、大小、颜色、光谱的红移等。本文的工作就是基于这些现有的参数完成的，而不是基于原始的图像和光谱。这些数据对于大尺度宇宙结构的研究有着重要意义。

当前，SDSS已经进入第二阶段观测（SDSS-II），该观测将持续到 2008 年 6 月，它包含 *the Sloan Legacy Survey* 、*SEGUE* 和 *the Sloan Supernova Survey* 3 个分支，从而为更好的解答宇宙的本质、星系与类星体的起源，以及我们所在的星系（银河系）的演化做好铺垫。关于SDSS巡天计划的详细介绍可参考SDSS的官方网站http://www.sdss.org[62]。

3.2.2 SDSS 星系参数

SDSS 的数据处理是通过一系列的自动处理软件包完成的。其中，天体的基本测光参数，如星等、颜色、轮廓、大小等主要是由 *Photo* 软件包完成的；而光谱的基本参数，如红移、光谱型等则由一维光谱和二维光谱处理软件（*idlspec1d* 和 *spectro2d*）给出的。关于软件包处理的具体过程及输出参数等的详细介绍可参考 Stoughton 等人关于 SDSS 早期释放数据的文章[60]。

在测光数据处理中，SDSS采用了一种修正过的*Petrosian*星等系统[63]。由于星系表面亮度会随着红移的增加而降低，传统的等亮度星等的定义会使得高红移星系的星等测量产生偏差。而SDSS采用的这种*Petrosian*星等系统对于给定某类星系，无论其红移高低，总是能测量其固定组份的流量。其具体过程如下，首先需要定义一个*Petrosian*半径r_p，该半径的定义为在该半径处局部表面亮度是其内部平均表面亮度的 20%，即：

$$\frac{\int_{0.8r_p}^{1.25r_p} 2\pi r I(r)dr/[\pi(1.25^2-0.8^2)r^2]}{\int_0^{r_p} 2\pi r I(r)dr/[\pi r^2]}=0.2 \qquad (3\text{-}2)$$

其中*I(r)*是表示星系在方位角平均后的半径方向上的表面亮度的轮廓。基于r_p，*Petrosian*流量（星等）F_p则定义为其两倍r_p内的总流量：

$$F_p=\int_0^{2r_p} 2\pi r I(r)dr \qquad (3\text{-}3)$$

根据这个定义，对于一个具有标准指数光度轮廓的星系来说，*Petrosian*星等包含了其总流量的98%；而对于*de Vaucouleurs*轮廓来说，该值仅为80%。值得注意的是星系的r_p由*r*波段的光度轮廓确定的，不同波段的*Petrosian*流量都同样积分到$2r_p$。这样可以保证不同波段得到的星等，即颜色没有孔径误差。

在*Petrosian*星等基础上，*Photo*还给出了另外一些重要的参数，如*PetU50*、*PetU90*、*PetG50*、*PetG90*、*PetR50*、*PetR90*、*PetI50*、*PetI90*、*PetZ50*、*PetZ90*，即分别包含在各波带（*u,g,r,i,z*）处50%和90%的*Petrosian*流量处半径。其中有两个最为重要的半径量*PetR50*和*PetR90*，基于这两个半径，星系的另一个基本参量致密度（concentration）*c*，被定义为*c=R90/R50*。该参数的实质是对星系光度轮廓的一种描述，因此与星系的形态有较强的相关性。

除了以上的*Petrosian*星等系统外，SDSS的标准输出参数还包括了模型星等。所谓模型星等就是指利用指数轮廓和*de Vaucouleurs*轮廓，并同时考虑星系的方位角以及长短轴比等参数对星系的光度轮廓进行χ^2拟合，并记录相应的拟合概率*exp_L*和*deV_L*，星系按指数轮廓拟合：

$$I(r)=I_0\exp(-1.68r/r_e) \qquad (3\text{-}4)$$

按*de Vaucouleurs*轮廓拟合：

$$I(r)=I_0\exp\{-7.67[(r/r_e)^{1/4}]\} \qquad (3\text{-}5)$$

每一种拟合都有它的优势，模型星等就是取这两种拟合中较好的一个，为最终得到的星等。为考虑消除星际消光带来的影响，本文我们也采用了进行红化改正后的红化星等，即模型星等减去消光量。

除了以上这些参数外，SDSS 给定的标准测光输出还包括 *PSF* 星等、星等的椭率、位置角、长短轴比等其他很多参数。但由于本文中没有涉及这些参数，所以这里不再一一介绍，一些基本的参数见下表 3-1，具体可参见关于 SDSS 早期释放数据的论文[60]。

如前所述，SDSS的光谱参数主要由*idlspec2d*和*spectro1d*分别处理，其中前者主要处理二维光谱，对其进行流量和波长的改正定标并获得一维光谱；而后者则处理一维光谱，测量光谱中的吸收线和发射线，并对光谱进行分类，确定红移等。由于本文主要关心星系的测光性质，在光谱参数中唯一使用的参量是星系的红移值，因此这些流程的细节这里不作详细介绍。SDSS中光谱红移测量的典型误差小于～10^{-4}。

表 3-1 SDSS 中有关星系参数的简单介绍

参数名称	所用符号	参数描述
modelMag	*model u,g,r,i,z*	模型星等
PetroMag	*Petro u,g,r,i,z*	*Petrosian* 星等
PetroRad	r_p	*r* 波段 *Petrosian* 半径
PetroR50	R_{50}	*r* 波段 50%*Petrosian* 流量半径
PetroR90	R_{90}	*r* 波段 90%*Petrosian* 流量半径
deVRad		*r* 波段 *de Vaucoulerus* 轮廓半径
expRad		*r* 波段指数轮廓半径
deVAB		*r* 波段 *de Vaucoulerus* 轮廓轴率
expAB		*r* 波段指数轮廓轴率
lnLdeV		*r* 波段 *de Vaucoulerus* 轮廓概率的对数
InLexp		*r* 波段指数轮廓概率的对数
z	*z*	光谱红移
zConf		红移置信度
eClass		星系光谱主分量分析的光谱类型
eCoeffs		*eClass coefficients*(系数)
Model_ug…iz	*u-g,…,i-z*	基于 *u,g,r,i,z* 波段模型颜色
Petro_ug…iz	*u-g,…,i-z*	基于 *u,g,r,i,z* 波段的 *Petrosian* 颜色
sb_r	μ_{R50}	半光表面亮度
Inv_CI	CI_{inv}	致密度指数的倒数 R_{50}/R_{90}
Absolute_model Counts	M_r	*r* 波段绝对模型星等
petroR90_kpc		*r* 波段绝对 *Petrosian* 半径
Mpa_stellarmass		恒星质量

3.2.3 样本的选择和参数提取

本工作所使用的样本是 SDSS 在 2004 年 3 月第二次释放的星系数据（SDSS Data Release 2），它覆盖了 3324 平方度，包含 367,360 天体光谱。样本的选择是按 2004 年 Vanzella 给出的标准[4]，即 *r* 波段的 *Petrosian* 星等大于 17.77，光谱红移置信度大于 0.95 并且没有警告标记（*zWarning*=0）。对 DR2 中满足条件的星系样本共有 159,346 个，其红移分布的统计直方图见图 3-2。将这些样本随机分为三部分：60,000 个、20,000 个和 79,346 个分别作为神经网络的训练样本、评价样本和测试样本。其中训练样本被用来构造网络模型，评价数据是为了防止出现“过拟合”现象而对模型进行监测，而测试样本则是对最终的网络模型作测试。这样由训练样本训练和评估样本评估而得到的模型或回归器给出测试样本的红移预测值，再与测试样本的实际光谱红移值比较，从而可以得出该网络模型的相对误差。

我们所用的参数主要是 SDSS 在各波段给出的不同形式的星等：*Petrosian* 星等、模型星等及红化校正星等、颜色及各波带（*u,g,r,i,z*）处 50%和 90%的 *Petrosian* 流量处半径。这些参数可以作为神经网络的输入参数进行实验。

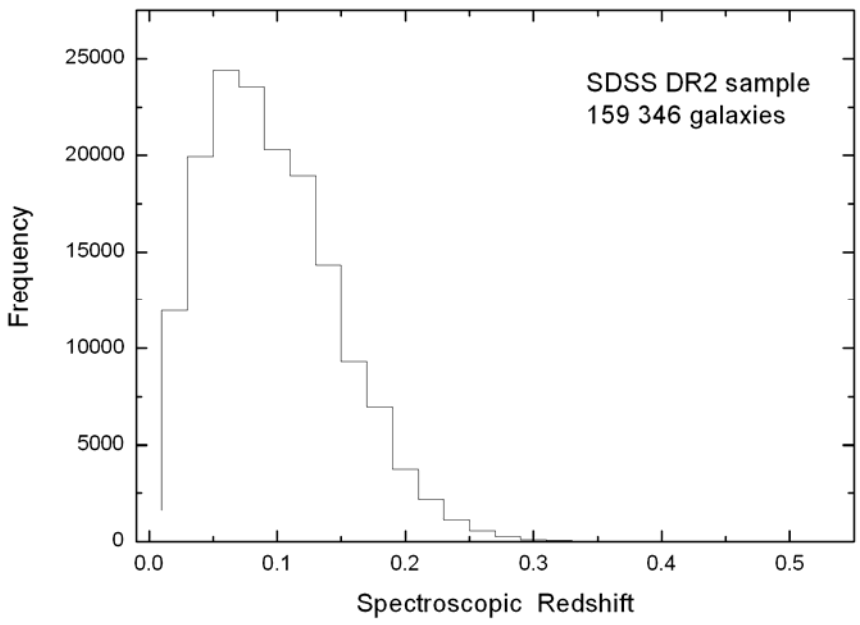

图 3-1　所选样本红移的统计直方图

3.3 应用神经网络预测测光红移

该实验使用的是Matlab工具箱中的*nnet*工具包中的前向多层神经网络与BP算法结合。实验中，训练数据和测试数据是相互独立的，又要求训练数据能代表测试数据的特性，因此我们从总样本中随机抽取训练和测试数据。用训练数据构造的网络模型可以成功地应用于测试数据，且具有较好的学习能力和泛化能力。为评价预测红移的精度，定义出神经网络输出红移值（*zNN*）和光谱红移值（*zspec*）之间的偏差（σ_z），如下式：

$$\sigma_z = \sqrt{\frac{1}{N}\sum_i (zNN_i - zspec_i)^2} \qquad (3\text{-}6)$$

其中 *N* 是星系样本数，*i=1,2,...,N*，此式给出由神经网络预测的测光红移的统计评估。

实验分别用 *Petrosian* 星等、模型星等和红化校正星等三套主要参数及相应的颜色指数与 *r* 波段星等的组合，我们这里只给出 *r* 波段的基本测光参数，是因为 *r* 波段是 SDSS 中进行测光处理的基准波段。实验中又增加了所有波带（*u,g,r,i,z*）处 50%和90%的 *Petrosian* 流量处半径作为输入参数，进行不同的模式输入的比较。下面分别对各组实验进行说明。

3.3.1 *Petrosian* 星等实验

此组实验中，首先使用五个不同波段的*Petrosian*星等（*Petrosian u,g,r,i,z*）作为神经网络的输入参数。隐层节点通过尝试和纠错过程来选择，这里没有使用Bayesian信息标准等定量方法，因为没有任何一个对大数据量选择隐层节点的程序会比尝试和纠错过程更有优势。在训练过程中，分别用一个或两个隐含层和不同的节点来尝试，最终储存具有最小训练误差的相应权值。当网络结构为 5:10:10:1（5 个输入节点、两个隐含层各有 10 个节点单元、1 个输出节点）时，实验结果最好，所得预测值和真实值之间的结果的最小偏差为σ_z=0.027031。

为了比较其它参数对神经网络测光红移的影响，我们改用*Petrosian*星等的颜色指数*u-g,g-r,r-i,i-z*及*r*波段*Petrosian*星等参数输入神经网络。同样取不同的网络结构及隐层节点进行尝试，选择出最好的权值分布，确定的网络结构为 5:10:10:1，应用于测试样本时的偏差σ_z=0.026717。可见，使用第二套参数比第一套的五个*Petrosian*星等要好。但是我们期望预测红移的精度越高越好，即所得的弥散度越小越好。因此我们尝试增加一些参数来考察红移的精度是否有所改进。

当然，参数越多包含的信息也相应地增加，更有利于了解有关星系红移的特性。为了增加信息，我们在第二套参数基础上增加了*r*波段处 50%和 90%的*Petrosian*流量处半径（*PetR50*、*PetR90*）。将 7 个参数输入神经网络进行实验尝试，最终选择的网络结构为 7:16:16:1,所得出的测试数据评估误差σ_z =0.025048。由此可见，新信息的加入使得结果改善，即红移精度有所提高。

通过上面的实验，可以看出在神经网络的输入参数中增加信息是一种提高网络的泛化能力的有效途径。那么再考虑加入其他波段 50%和 90%的*Petrosian*流量处半径（*PetU50*、*PetU90*、*PetG50*、*PetG90*、*PetI50*、*PetI90*、*PetZ50*、*PetZ90*）和各波段星等信息，测光红移的结果能否得到提高？为此我们做了以下实验，将所提到的 19 个参数输入神经网络，选择了 20 个神经元的单隐含层。没有使用更复杂的网络结构是因为通过多次实验显示在神经网络结构中增加更多的隐含层和节点数并没有使结果有更大的改善。最终确定的网络结构是 19:20:1，结果偏差降到σ_z =0.021596。为更清楚起见，图 3-2 给出了使用 19 个输入参数的神经网络预测红移结果与SDSS光谱红移值比较，由该图可以看出预测值与实际值之间符合的相当好，数据弥散度很小，基本上分布于对角线附近。为比较各种参数对红移精度的影响，表 3-2 列出了每套参数预测测光红移的偏差。

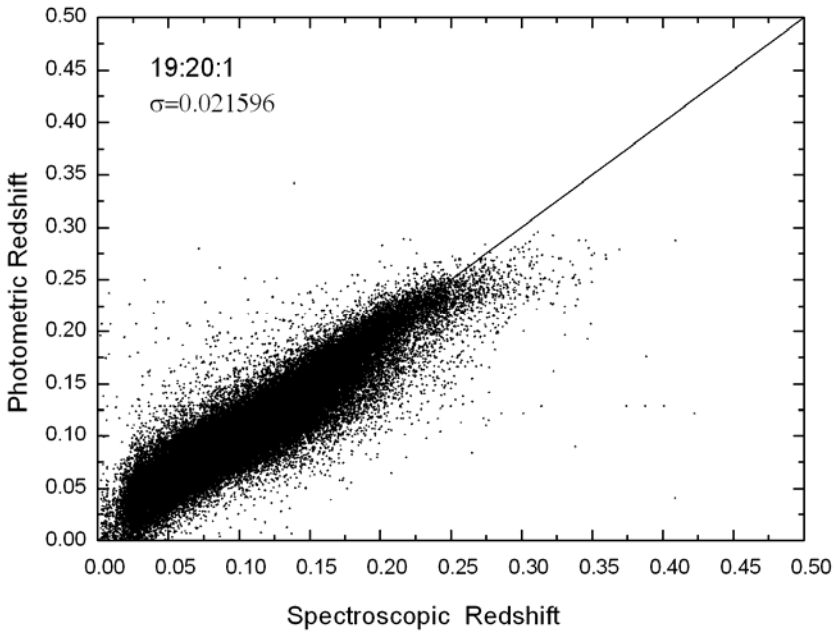

图 3-2　使用 Petrosian 星等的 19 个参数的神经网络

测光红移与 SDSS 光谱红移的比较

表 3-2　各输入参数对训练样本和测试样本预测红移的结果弥散

参数个数	输入参数	训练数据偏差 σ	测试数据偏差 σ
5	*Petrosian u, g, r, i, z*	0.026939	0.027031
5	*Petrosian u-g, g-r, r-i, i-z, r*	0.026535	0.026717
7	*Petrosian u-g, g-r, r-i, i-z, r,PetR50,PetR90*	0.025002	0.025131
19	*Petrosian u-g, g-r, r-i, i-z, u, g, r, i, z, PetU50,PetU90,PetG50,PetG90,PetR50, PetR90,PetI50,PetI90,PetZ50,PetZ90*	0.021502	0.021596

3.3.2 *Model* 星等实验

SDSS 给出的各种星等参数，那么究竟使用哪些参数才能更有效地预测测光红移呢？为此，我们又以模型星等作为主要参数进行了实验，以获取对测光红移来说最为有效的参数。

同样先用五个不同波段（*u,g,r,i,z*）的模型（*Model*）星等作

为网络的输入参数，随机初始化权值分布，训练选择网络，为了防止网络出现“过拟合”现象，使用评价数据监督使其中期停止，最终网络结构是 5:12:8:1，反馈训练 80 次。使用模型星等参数训练的网络对测试数据产生的偏差σ_z=0.0233。取代五个星等，使用模型星等的色指数（即颜色）*u-g,g-r,r-i,i-z*及*r*波段*Model*星等作为网络输入，实验得出与上面相同的网络结构 5:12:8:1，但是不同的初始化权值分布，最终的网络预测红移误差σ_z =0.0221。

作为参数的比较，同样在上面的五个参数基础上增加了*r*波段处 50%和 90%的*Petrosian*流量处半径（*PetR50*、*PetR90*）信息，初始化网络的最大训练次数是 3000 次，实验中改变初始化权值分布并采用初期终止避免网络的“过拟合”现象。使用多种网络结构训练以选择最好的、简单的形式，最终确定为 7:12:8:1，该网络预测红移的误差为σ_z =0.02075。可以看出，这个结果较上一实验有很大的提高，可以与其他文献中使用神经网络测光红移相比较[64-67]。可见，*PetR50* 和*PetR90* 的应用对提高网络性能，改善测光和光谱红移的一致性上很有帮助。

最后，相应地增加其它波段处 50%和 90%的*Petrosian*流量处半径（*PetU50*、*PetU90*、*PetG50*、*PetG90*、*PetI50*、*PetI90*、*PetZ50*、*PetZ90*）和其它波段星等信息，通过实验进一步检验能否减小结果偏差。19 个参数作为神经网络输入，不断尝试选取网络结构和初始权值，实验确定的网络结构为 19:12:8:1，得出误差分布是σ_z=0.020465。可见参数的增加再次改善了网络的性能，提高了测光红移的预测精度。将 19 个参数对 79,346 个星系测试样本的红移评估与光谱红移的比较，如图 3-3 所示。关于使用模型星等各套参数预测测光红移的结果参看表 3-3。

表 3-3　使用模型星等各套参数测光红移结果比较

参数个数	输入参数	训练数据偏差 σ	测试数据偏差 σ
5	*Model u, g, r, i, z*	0.023354	0.023321
5	*Model u-g, g-r, r-i, i-z, r*	0.022006	0.022097
7	*Model u-g,, g-r, r-i, i-z, r, PetR50,PetR90*	0.020765	0.02075
19	*Model u-g ,g-r, r-i, i-z, u, g, r, i, z，PetU50,PetU90,PetG50,PetG90,PetR50, PetR90,PetI50,PetI90,PetZ50,PetZ90*	0.02034	0.020465

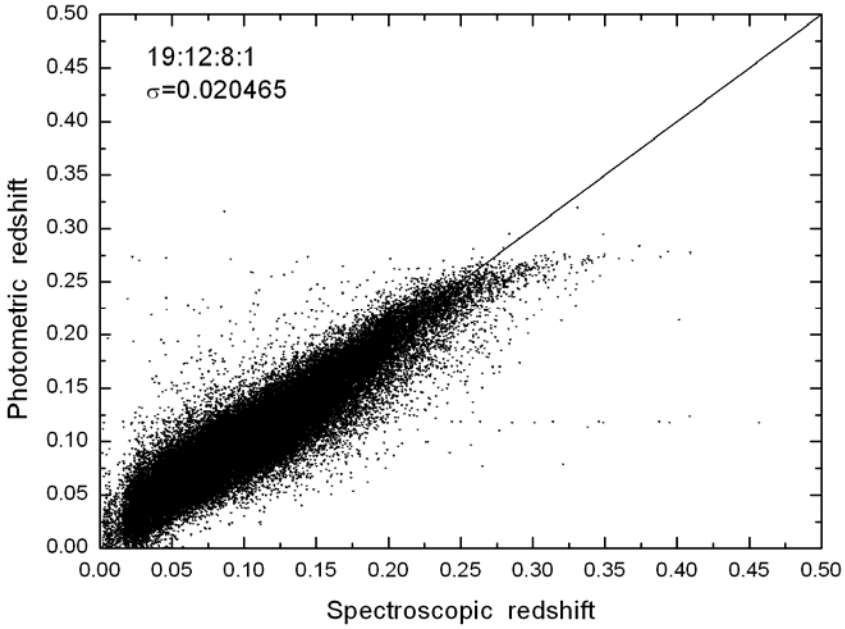

图 3-3　使用 19 个参数的模型星等测光红移与光谱红移比较

3.3.3 *Dereddening* 星等实验

类似于*Petrosian*星等和模型星等测光红移实验，我们也讨论了红化校正星等（*Dereddening magnitude*）作为神经网络输入参数估计测光红移。具体如下：使用与前两实验相同的样本数据，

设定训练的最大步数为 3000 步，尝试有不同节点数的一层或两层隐含层的网络结构，给予不同的初始化权值分布。对各种不同的参数，对应的最终网络结构分别为 5:5:5:1、5:5:10:1、7:10:1 和 19:12:1，它们相应的红移偏差均被列在表 3-4 中，可见最好的结果是 19 个输入参数的网络，其结果偏差为σ_z=0.020184。19 个红化校正星等参数预测的红移与光谱红移值比较见图 3-4。

3-4 使用红化校正星等各套参数测光红移结果列表

参数个数	输入参数	训练数据偏差 σ	测试数据偏差 σ
5	*dereddening u, g, r, i, z*	0.021371	0.02388
5	*dereddening u-g, g-r, r-i, i-z, r*	0.021081	0.021097
7	*dereddening u-g, g-r, r-i, i-z, r PetR50,PetR90*	0.020821	0.020689
19	*dereddening u-g, g-r, r-i, i-z, u, g, r, i, z, PetU50,PetU90,PetG50,PetG90,PetR50, PetR90,PetI50,PetI90,PetZ50,PetZ90*	0.020174	0.020184

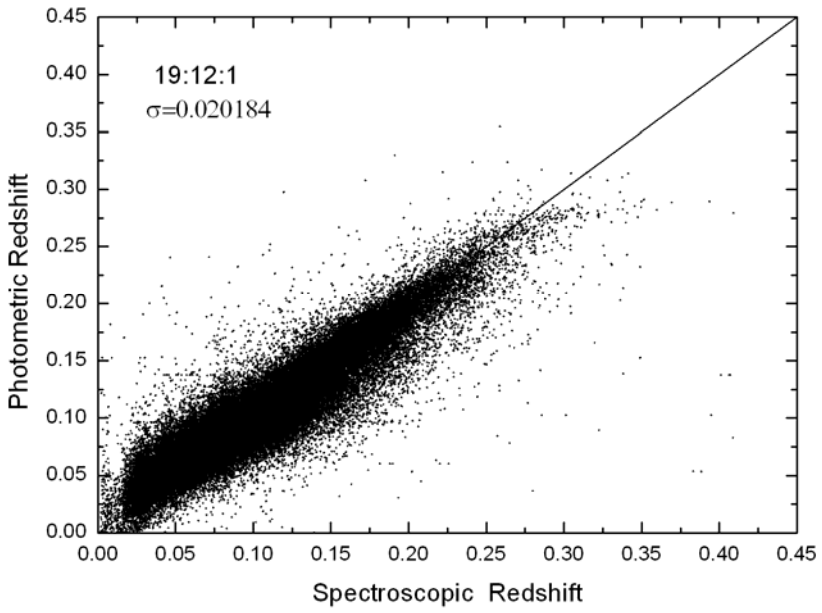

图 3-4 使用 19 个参数的红化校正星等测光红移与光谱红移比较

3.3.4 结果比较与小结

本文我们用前向神经网络对大量的参数作测光红移实验，发现了一些有意义的结论。实验结果表明，不论使用*Petrosian*星等、模型星等还是红化校正星等，都有一个共同的特性：所使用的输入参数越多，神经网络的预测精度也就越高。随着训练网输入参数的增加，将有更多的可以提高网络预测和泛化能力的有用信息，使得最终网络预测的精度也有所提高。而且可以看到，对于相同的数据和参数结构来说，*Petrosian*星等、模型星等和红化校正星等都可以作为神经网络的输入参数来估计测光红移，但是红化校正星等要比*Petrosian*星等、模型星等更好，更有效。比较另外两组实验可见模型星等又比*Petrosian*星等更有优势。对我们实验的测光红移结果最小的偏差σ_z=0.020184，对覆盖整个天区样本的统计研究工作，这样的精度是可用的，可以帮助大尺度结构研究者更容易地探讨一些与宇宙学相关的课题。

随着天文观测的深入发展，已经有越来越多的可用参数。测光红移面临的主要问题是与红移没有明显关系的不合适参数导致较大的弥散和误差，因此选择合适的、有效的参数让天文学家方便使用将是一个更为迫切的问题，同时又是一个严峻的挑战。我们根据实验结果，选择消光后的红化校正星等及*PetR50* 和*PetR90* 参数组合，大大提高了预测精度，从而为测光红移的研究者提供了一套较为优化的参数组合，解决了测光红移参数难以选择的问题。当然，为了改善测光红移的精度，我们将要考虑更多波段的参数，像2MASS中的J,H,K_s波段。另外随着参数的增加，必然会造成训练网络的速度减慢，我们将进一步考虑特征提取（如主成分分析方法PCA）方法在多维参数空间中提取主要成分，揭示其潜在的因素或信息，从而提高神经网络的效率。

这里只是将神经网络应用于测光红移之中，其实它在天文中有广泛的应用，主要是因为它在处理多参数的非线性问题时更有用、更高效。目前神经网络已应用到天文的各个领域中（可参看本论文的 2.4 节），如恒星光谱分类[68-72] 、星系光谱分类[73-74]、星系形态分类[75-77]和在宽视场图片区分恒星和星系[78-80]等。尽管人工神经网络在某些问题领域有很大的优势，但

它也存在一些缺陷：首先神经网络训练时间较长；其次，它是在参数空间进行局域搜索，因此常限入局域极小值；再者一旦被训练，神经网络是一个黑箱，在某种意义上说对于使用者很难解释它的学习规则。因此要想获得学习规则，可以考虑用决策树或决策规则方法。而我们在用神经网络处理问题时，需要掌握它的几个关键要素：即网络模型的选取、网络结构设计、数据的预处理、训练模式与测试模式的生成、实验与参数调整及后续处理等。

第四章　基于 K 近邻算法的天体自动分类

4.1　天体的自动分类

面对海量的天文数据，我们将面临许多实质性的挑战：如何有效地分析和利用这些数据，挖掘出隐藏在数据中的信息和知识？首先就是要将包含着复杂信息的数据进行分类，即根据数据内在的相似性把它们分为几个类群，它也是目前巡天工作的首要任务。

天文学从始至终都包含着分类问题，传统的分类方法是使用一些人机交互式的天文软件，对人的经验要求很高，而且当光谱的信噪比低时，人工处理的难度也大大增加。随着现代的多波段数字巡天项目的开展，如象 LAMOST 这样的产生海量数据的巡天项目，靠人工交互逐个完成所有光谱或测光的分类是不可能的。为了解决这一难题，应该构造一个完全自动的系统，尽可能使用一些智能化的工具，以快速准确的完成天体的自动分类，而这一切必须依靠强有力的分析技术支持才能完成。

分类是数据挖掘、机器学习和模式识别中一个重要的研究领域。分类的目的是：分析输入数据，通过在训练集中的数据表现出来的特性，为每一个类找到一种准确的描述或者模型。这种描述常常用谓词表示。由此生成的类描述用来对未来的测试数据进行分类。尽管这些未来的测试数据的类标签是未知的，我们仍可以由此预测这些新数据所属的类。而这只是预测，不能完全肯定。我们也可以由此对数据中的每一个类有更好的理解。也就是说：我们获得了对这个类的知识。

进行分类，首先要构造分类器。构造分类器的方法有多种，如统计方法、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。

一般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简单实用，而神经网络过程黑箱操作，产生的结果则让人难以理解[1]。

另外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，这都或多或少的对分类效果产生影响。普遍认为不存在可以适用于所有特点数据的分类方法。

目前典型的数据挖掘方法已被应用到了天文学中，概括起来主要有：

（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[81-83]，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[84-87]，如 EM(Expectation Maximization)，MCCV(MonteCarlo 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）[88-90]，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。

本文采用一种非参数法的方法或基于范例的学习方法—*K* 近邻算法对天文数据分类，目的是将类星体从恒星和星系的样本中分离出来。

4.2 *K* 近邻算法原理

K 近邻算法（*k*-Nearest Neighbor algorithm，简称 *k*NN）是最简单、最常用的机器学习方法之一。它是一种重要的非参数分类方法，具有灵活、有效的特点，已在许多领域得到了广泛的应用。同时它又是一种基于实例的分类算法，训练阶段把已知类型的训练实例简单地存储起来，以备分类时查询使用。每当新的待分类实例到达时，它就检索训练实例空间，找到 *k* 个与查询实例最相似的训练实例（*k* 个最近的邻居），然后将 *k* 个邻居的类标号中最普遍的类别作为新实例的类标号，即得出新样本的预测值。

首先构造分类器，这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。如将训练样本$(x, f(x))$存储于内存空间中，x是由n维特征向量$a_1, a_2, a_3 \ldots a_n$组成的样本实例，$f(x)$是相应的类别标注。对每个训练实例$(x, f(x))$，将它加入训练实例表*training_examples*中。给定一个待分类样本x_q，计算它与各训练样本之间的距离，最常使用的是欧式距离（Euclidean distance）：

$$d(x_q, x_i) = \sqrt{\sum_{r=1}^{n} (a_r(x_q) - a_r(x_i))^2} \qquad (4\text{-}1)$$

其中 a_r 是 x 的第 r 个特征向量。在训练实例表*training_examples*中找出与测试样本x_q距离最近的k个邻居，将k个邻居的类标号中最普遍的类别作为待测实例的类标号，即分类完成[91]。

由于 *k* 近邻在分类前并不作任何处理，它把工作都推迟到待

分类实例到达时来做，因此不计时间复杂度。在分类阶段，它要顺序扫描一遍训练集以确定 k 个与查询实例最近的邻居，然后确定其中最普遍的类标号，所以分类一个文档的时间复杂度为 $O(N)$。

最近邻算法原理简单，使用也十分方便，但也具有明显的缺点，比如计算量大，存储量大等，它需要存储全部训练样本，以及繁重的距离计算量。

4.3 样本和参数的选择

天体辐射的能量分布覆盖了非常宽的波段，从射电到红外、光学、紫外、X 射线甚至到 γ 射线，每个波段上的观测都带来了有关天体性质的重要信息。同样一个天体在不同波段上的表现是可以完全不同的，如蟹状星云的光学图像显示出了电离氢的分布，射电图像显示了中性氢的分布，红外图像显示了尘埃和分子云的分布，而 X 射线图像显示了高温（千万度）热气体的分布和其中存在的中子星。因此要研究天体的物理过程，就必须结合几个波段上的数据来一起进行分析。天文数据是全波段的，同时又是高度相关的，因此在高维参数空间内进行天文研究具有非常重要意义的。

一般来说，天文学上将活动星系核（Active Galactic Nucleus, AGN）作为活动天体，而将恒星和正常星系视为非活动天体。从恒星和星系中分离出有特殊性质的活动星系核是一项粗分类任务，但它可以为天文学家提供一些活动星系核的预选源。本文使用张彦霞等人通过交叉证认确定的已知类型样本作为训练集[92]，使用 k 近邻算法对未知类型的天体进行分类。我们知道活动星系核是有着特殊性质的一类天体，它的特性不能仅从一个波段表现出来，因此这样的分类工作需要考虑来自天体的多个波段信息。本文中我们主要使用了三个波段的观测星表即 ROSAT 亮源表与弱源表（X 射线）、USNO-A2.0（可见光）和 2MASS（近红外）星表，下面对这几个星表作简单介绍：

ROSAT All-Sky Survey (RASS)是德国的 X 射线天文台在

1990 年 6 月 1 日开始启用的全天巡天星表，它包含 ROSAT 的全天巡天亮源表和弱源表。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数据和洞察力。

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$J(1.25\mu m)$、$H(1.65\mu m)$和K_s $(2.17\mu m)$，它们的不确定性分别约为 0.001 个星等。

对于活动星系核，采用 Veron （2000）的活动星系核表[93]，该表中包含了 13,214 个类星体、462 个 BL Lac 天体和 4428 个活动星系（其中 1711 个为 Seyfert1）。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[94]。SIMBAD 星表是由斯特拉斯堡 CDS 创建和维护的，其收集了大量的约 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。

张彦霞等人的工作中用到的数据样本[92]，是由这三个数据库中的资源通过位置交叉证认获得的，其过程如下：

（1） 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。

（2）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 $J(1.25\ \mu\ m)$、$H(1.65\ \mu\ m)$,和$K_s(2.17\ \mu\ m)$星等。

（3）最后，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（2）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 5547 个。

本文对该工作进行了后续研究，选取来自不同波段数据的参数为*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）。这样的十个参数携带着天体在三个波段的不同表现信息，可以根据它们来研究天体的内在属性，将活动星系核从恒星和星系中分离出来。

4.4 *K* 近邻分类实验及结果

*K*近邻算法属于监督学习算法，需要构造分类器。因此我们将已选好的 5547 个样本随机分成两部分，一部分 2774 个作为训练集用于构造分类器，余下的 2773 个数据则作为测试样本检验分类器所获得的分类精度和效率。如果我们构造的分类器效果是好的，那么就可以用它来对其它未知样本作类别预测。首先要给出训练集样本的类别，为了简化，我们将活动星系核（AGN）的标注设为-1，而将恒星和正常星系的标注设为 1。然后将选取的多波段参数（*B-R*、*B+2.5log(CR)*、*CR*、*HR1*、*HR2*、*ext*、*extl*、*J-H*、$H-K_s$、*J+2.5log(CR)*）作为*K*近邻算法的输入参数进行分类实验。在实验中*K*的选取是通过尝试得到的，先给定*K*一个小的整数，依次逐渐增加，记录各不同*K*值所得的分类精度和构建分类器所需的时间，通过比较分类精度的好坏来评价各不同*K*值的分类效果。在本实验中我们选取的*K*值是从 2 到 17，*K*取不同值时分类器的分类精度和运行时间比较结果如下表 4-1。

表 4-1　对不同 K 值分类器的分类精度和运行时间的比较

K 值	运行时间（秒）	AGN 误分个数	恒星、星系误分个数	分类精度
2	5.568	30	50	97.12%
3	5.569	25	40	97.66%
4	5.698	23	43	97.62%
5	5.738	27	41	97.55%
7	5.829	26	43	97.51%
8	5.878	24	44	97.55%
9	5.908	23	42	97.66%
10	5.948	23	40	97.73%
11	6.049	22	44	97.62%
12	6.099	22	41	97.73%
13	6.129	24	44	97.55%
15	6.189	25	46	97.44%
17	6.309	26	44	97.48%

表中包含有 K 的不同取值、构建分类器和对测试样本预测所用的运行时间、对测试样本中 AGN 的误分类个数、对恒星和星系的误分类个数及 K 近邻算法对测试样本分类的精度。从此表可以看出，当 K=10 和 K=12 时，K 近邻算法的分类精度最高为 97.73%，这个结果已经相当可观。但是比较两者的运行时间，明显地 K=12 比 K=10 的分类器耗时要长。另外不论 K 取值如何，K 近邻算法对测试样本的分类精度都在 97%以上，构造分类器和运行测试样本所用的时间仅在几秒之内，可见 K 近邻算法分类效率很高，分类效果显著，非常适合于对天文数据的分类实验。

为了更清楚地显示实验结果，图 4-1 给出了 K 的取值与测试样本分类精度的对照。由图可以看出，在 K 近邻分类算法中分类精度不是随着 K 值的增加不断提高的，精度曲线不是平滑递增而是迂回曲折的。从理论上说，较大的 K 值应该能够减小训练样本的噪声而使精度曲线更平滑；但根据实践经验，一般 K 的取值应为几或几十这样小的整数而不是成百或上千那样的大数。为了进一步证实这点，这里我们也用了较大的 K 值做实验，如 K=30 和

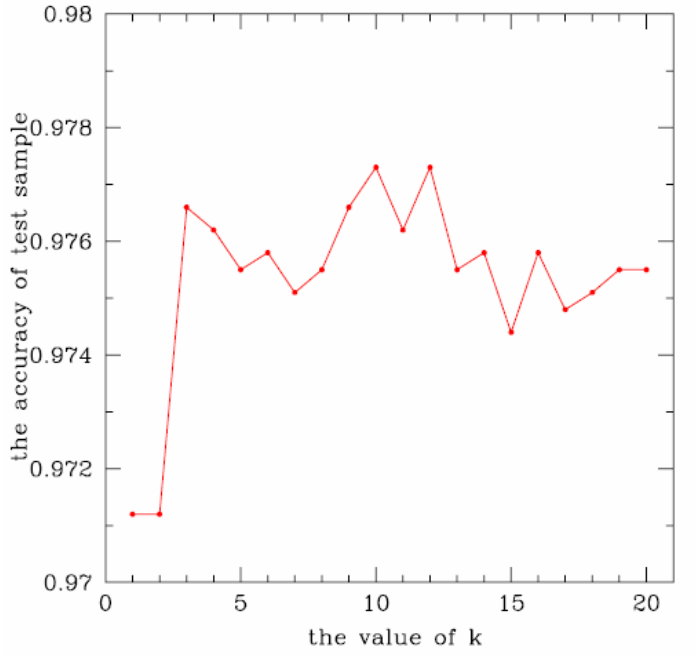

图 4-1　对不同的 K 值所得的预测精度

K=50 的分类精度分别是 97.44%和 97.48%。相比较而言，它们的精度并没有得到提高但所用的运行时间却更长了（分别为 7.20 秒和 8.262 秒）。所以综合比较 K 取不同值时，分类器的精度和其运行时间，最终确定了 K=10，以此分类器从恒星和星系多波段数据中分离出类星体的精度达到 97.73%。

4.5 本章小结

本章我们使用了 K 近邻算法来构造自动分类器，通过尝试选取不同的 K 值，比较分类精度和构造分类器及对测试样本预测所用的时间最终确定 K 的取值为 10。从实验结果表示将 K 近邻算法应用于较大的数据库中能够产生较高的分类精度并且它的运行时间相当短，效率比较高。同时 K 近邻算法具有较强的泛化能力，应用潜能很广。

K 近邻方法是数据挖掘分类算法中的一种灵活、有效的方法，在天文上可以用于对天体分类，从而可以在大数据库中预选观测源，可以服务于 LAMOST 项目的输入星表。对于 K 近邻算

法分类来说，各种类型的数据（包含星系的形态、测光和光谱等等）都可以用来训练获得分类器，对进行天体分类及预选感兴趣的天体等。但是当缺少训练样本时，我们可以使用一些非监督方法或离群数据发现（outlier finding）的方法来找出不寻常的、稀有的或新类型的天体或天文现象。随着虚拟天文台的发展，*K* 近邻方法将集成到国际虚拟天文台的数据挖掘的工具箱中。

第五章　星系形态分类及测光红移

5.1 哈勃星系形态分类

星系在还没有被认证的时候被认为是星云，如旋涡星云。还有那些有自己特殊名字的星系，如仙女座大星云、猎犬座大星云、大小麦哲伦云等。那时之所以把星系认为成星云，是它们具有与银河星云相类似的形态，都呈云雾状，而且不像星团那样可以被分解开。

哈勃对仙女座星云距离的开创性研究为研究河外星系打开了大门。现在我们知道，在可观测宇宙中约有 10^{12} 个星系，这些星系中有许多星系的质量可与银河系相比拟；多数比银河系小；少数比银河系大得多。虽然在古代星系已被分类，但它们的类型并非是对星系完全客观的描述而是带有明显的人为色彩。几十年来，人们已经建立了几种分类系统，其中影响最大、应用最广的是哈勃分类法。

哈勃是星系天文学的开拓者，他依据大量的观测资料于 1926 年提出了一个星系的分类方案，后经过几十年，的修改和补充而逐步完善，世称哈勃分类。哈勃按形态分为如下四类[95]：

(1) 椭圆星系(Ellipticals)：E，外形呈正圆形或椭圆形，中心亮，边缘渐暗。按外形又分为 E0 到 E7 八种次型；

(2) 正常漩涡星系(Ordinary Spirals)：S 或 SA，外形呈旋涡结构，有明显的核心，核心呈透镜形，核心球外是一个薄薄的圆盘，有几条旋臂。正常漩涡星系又分为 a、b、c 三种次型：Sa 型中心区大，稀疏地分布着紧卷旋臂；Sb 型中心区较小，旋臂较大并较开展；Sc 型中心区为小亮核，旋臂大而松弛；

(3) 棒旋星系(Barred Spirals)：SB，在旋涡星系中有一类的核心不是球形，而是棒状，旋臂从棒的两端生出。棒旋星系可分为三类：①正常棒旋星系 SBa、SBb 和 SBc；②透镜型棒旋星系 SB0；③不规则棒旋星系 SBd 和 SBm；

(4) 不规则星系(Irregulars)：外形没有明显的核心和旋臂，

看不出旋转的对称性结构，呈不规则的形状。

漩涡星系（正常漩涡和棒旋）和椭圆星系包含了可观测宇宙的大部分质量，不规则星系仅占宇宙质量的百分之几。哈勃曾把更接近椭圆星系的星系称为透镜星系。透镜星系有明亮的核球和扁盘，但没有旋臂，形似透镜，以 S0 表示，并认为 S0 是椭圆星系与漩涡星系之间的过渡星系，他绘制了一幅形如音叉的图，如图 5-1 所示。

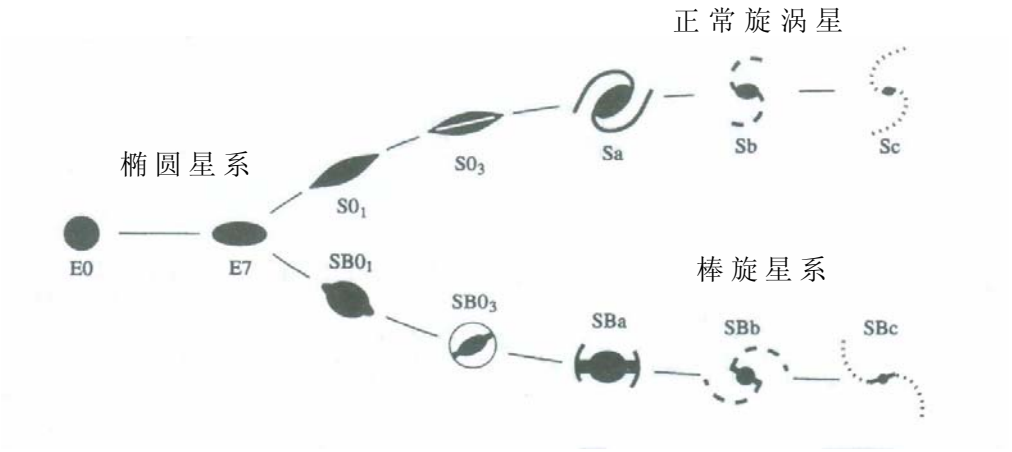

图 5-1 哈勃“音叉”图

哈勃提出这样一个演化序列，他认为椭圆星系会渐渐的演化成旋涡星系，即星系由左边的形态逐渐演化为右边的形态。这种令人怀疑的推测仍然有着生命力，习惯上将偏向序列左边起点处的星系看作是“早型星系”（E、S0 和 Sa），而偏向序列右边末端的星系被看作是“晚型星系”（Sb、Sc 和 Irr）。不过“早型”与“晚型”是相对的，例如旋涡星系对椭圆星系是晚型，但对不规则星系来说却是早型。

早型星系和晚型星系相比，有一些外在属性上的差异，一般来说，星系团中的早型星系用于恒星形成的大量气体都已经耗完或者被剥离，因此大部分早型星系都已经停止了恒星形成活动。而晚型星系越靠近星系团中心越表现出活跃的恒星形成活动。不同形态类型的星系在不同星等处对光度函数的贡献也是不同的。在亮星等处，在星系团中大部分星系是巨椭圆的早型星系，而在场中它们是巨晚型星系。袁启荣等人运用 BATC（全称

Beijing-Arizona-Taipei-Connecticut）多色测光体系分别研究了早型星系和晚型星系的颜色-星等关系[96]。得出结论只有早型星系的颜色和星等间是有一定的相关性的，星等越小星系越红，亮的星系要比暗的星系红。晚型星系则只能说明它们大部分是一些暗的蓝色星系，而它们的颜色和星等间没有明显的相关性。可见早型星系和晚型星系表现出不同的外在属性，我们有必要研究星系的形态特性，并采用自动算法对星系形态进行分类。

5.2 使用 *k-means* 算法自动聚类

5.2.1 *k-means* 算法原理

k-means 算法是一种基于样本间相似性度量的间接自动聚类方法，属于非监督学习方法。此算法以 k 为参数，把 n 个对象分为 k 个簇，以使簇内具有较高的相似度，而且簇间的相似度较低。相似度的计算根据一个簇中对象的平均值（被看作簇的重心）来进行。此算法首先随机选择 k 个对象，每个对象代表一个聚类的质心。对于其余的每一个对象，根据该对象与各聚类质心之间的距离，把它分配到与之最相似的聚类中。聚类相似度是利用各聚类中对象的均值所获得一个“中心对象”（引力中心）来进行计算的。其具体算法如下[1]：

输入：聚类个数 k，以及包含 n 个数据对象的数据库。

输出：满足方差最小标准的 k 个聚类。

处理流程：

（1）从 n 个数据对象任意选择 k 个对象作为初始聚类中心；

（2）循环（3）到（4）直到每个聚类不再发生变化为止；

（3）根据每个聚类对象的均值（中心对象），计算每个对象与这些中心对象的距离；并根据最小距离重新对相应对象进行划分；

（4）重新计算每个（有变化）聚类的均值（中心对象）。

k-means 算法是一种较典型的逐点修改迭代的动态聚类算法，其要点是以误差平方和为准则函数。逐点修改类中心：一个象元样本按某一原则，归属于某一组类后，就要重新计算这个组

类的均值，并且以新的均值作为凝聚中心点进行下一次象元素聚类；逐批修改类中心：在全部象元样本按某一组的类中心分类之后，再计算修改各类的均值，作为下一次分类的凝聚中心点。最终确定的 k 个聚类具有以下特点：各聚类本身尽可能地紧凑，而各聚类之间尽可能地分开。

5.2.2 *k-means* 算法自动聚类早型与晚型星系

对星系进行形态上的分类，目前最准确的莫过于人眼对图像的直接观察。在 SDSS 中，对于部分亮的星系，已经有这方面的尝试[97-98]。这种分类方法是用肉眼检查 SDSS 图像，根据星系的形态特征（如旋臂、棒、环结构以及尘埃带等）来进行分类。事实上这样的眼球分类是一件非常冗时的工作，而且也不大可能对 SDSS 这种大规模的数字巡天中的所有星系进行，对于大部分暗的星系就很难判断其形态了。为此，我们研究一种取代人眼的方法，即利用 SDSS 释放的所有星系数据进行大样本统计，根据星系的内在特性，采用一种非监督的 *k-means* 聚类算法将早型星系和晚型星系按形态自动分类。

研究已经表明，一些简单的测光参数或光谱参数和星系的形态有紧密的相关性，因此可以作为星系形态分类的简单判据。比如，Shimasaku等人的研究表明，星系的致密度参数c可以用来将星系分为早型（E/S0）和晚型（Sa/b/c,Irr）两大类[98]。利用眼球分类的样本，Nakamura等人进一步确证了 c=2.86 可将星系在S0/Sa处分开为早型和晚型两类，并且各自的完备性达到 80%以上[97]。根据星系颜色参数来分类是另一举创，Strateva等人对SDSS的早期数据的研究表明利用 u-r=2.22 这样的标准可以将星系分为早型（E/S0/Sa）和晚型（Sb/Sc/Sd）两类[99]。Blanton等人的研究表明，星系在 $(g-r)^{0.1}$~0.7 处可被分为各自具有明显不同性质的两组[100]。

尽管各种方法看上去都可以对星系进行形态分类，但是必须指出的是所有这些简单的分类准则都有很大的不确定性，而且只有统计上才显示其意义。比如说所有基于光谱进行的分类都可能有光谱的有限孔径（3 个角秒）带来的系统偏差，而所有基于颜

色和光度轮廓进行的分类则可能受到尘埃消光的影响。由于这些局限性，我们在本文中将不对星系作形态上的细分，而是应用 *k-means* 根据数据自身的特点将星系自动聚类为早型和晚型两大类。

实验中，我们提取 SDSS 巡天数据中的所有星系样本 582,512 个。首先进行样本数据的清理，消除或减少离群数据，去掉那些有缺值的离散数据。该过程可以通过人工操作或使用某些算法来实现。我们知道，*k-means* 算法具有自动聚类功能，因此可以用它对样本进行预处理。具体做法如下：首先选出有 *Petrosian* 星等缺值的数据，因此将五个波段的 *Petrosian* 星等作为 *k-means* 算法的输入参数，使该算法的程序运行，将自动聚出 219 个有缺值的样本，它们不符合实验的要求，故排除掉。再选择 *model* 星等缺值的样本数据，同样将五个波段的 *model* 星等作为输入参数进行 *k-means* 聚类，挑出 30 个有缺值的样本，将其去掉。最后挑选有关星系颜色的离群数据，将四个颜色指标(*u-g,g-r,r-i,i-z*)作为 *k-means* 算法的输入，自动聚类后得到 6 个离群数据。经过样本数据的清理和预处理后，得到的可用样本为 582, 257 个。

类似 Strateva 用颜色对星系分类的做法[99]，我们也是基于颜色进行分类的。在 Strateva 的文章中提到 *u-r* 是一个很有用的颜色指标，因此我们也把它作为输入参量，即五个颜色指标(*u-g,g-r,r-i,i-z,u-r*)作为算法的输入参数。*k-means* 聚类算法会根据所有星系的颜色特性将其自动聚为两类：即 300, 903 个和 281, 354 个。从每一组中随机抽取 50 至 100 个样本，按照位置信息与 NED 中所给星系类型进行比较，可知 300, 903 个为早型星系，而 281, 354 个为晚型星系。为了清楚起见，做出两类星系颜色 *u-r* 的统计直方图，如图 5-2 和图 5-3 所示，而总星系样本颜色的统计直方图见图 5-4。从 5-2 和 5-3 两图中，我们可以看出，使用 *k-means* 自动聚类算法大约在 *u-r*=2.65 处会将总的星系样本分为早型星系和晚型星系两类。

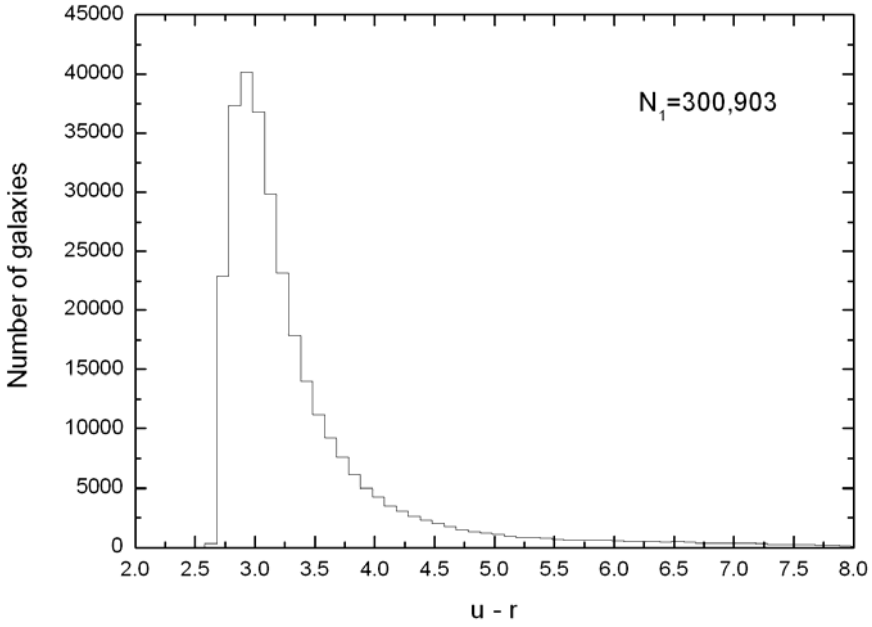

图 5-2　早型星系 *u-r* 颜色的统计直方图

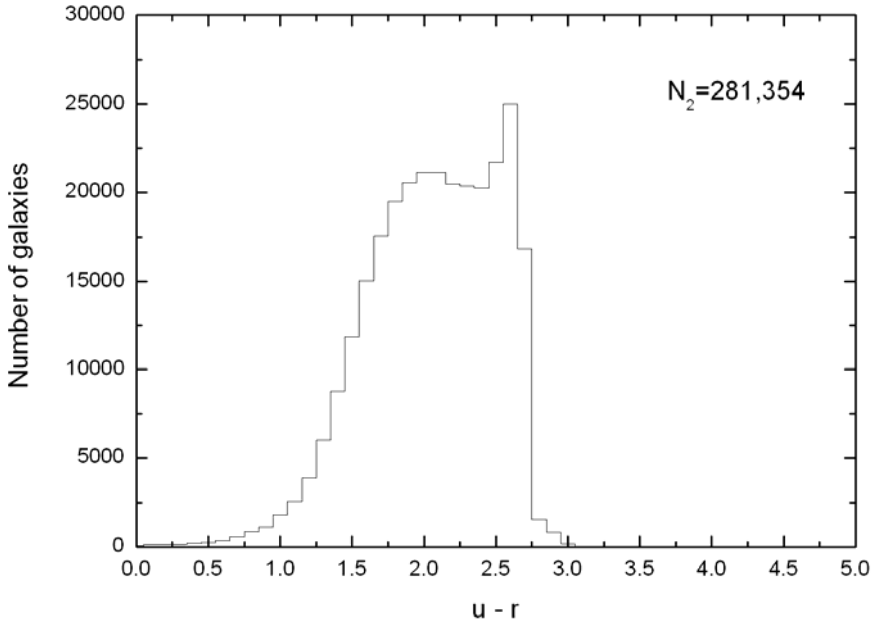

图 5-3　晚型星系 *u-r* 颜色的统计直方图

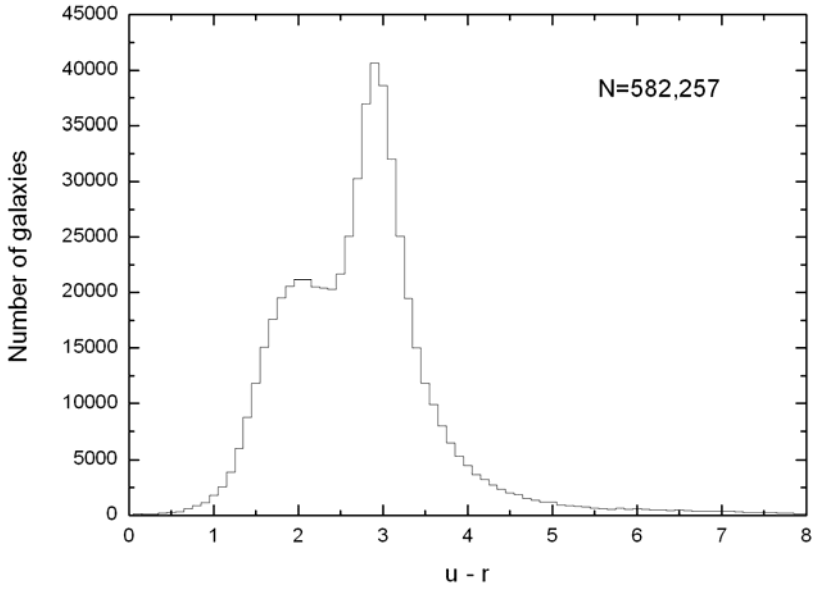

图 5-4　SDSS 所有星系样本 *u-r* 颜色统计图

将两类星系数据分别描绘在双色图(*g-r*，*u-g*)上，如图 5-5 所示，其中红色数据点为早型星系，黑色数据点为晚型星系。二者之间的分界线很明显，与 Strateva 文献中的图 4 相符，说明使用 *k-means* 自动聚类算法所获分类结果是比较合理的。

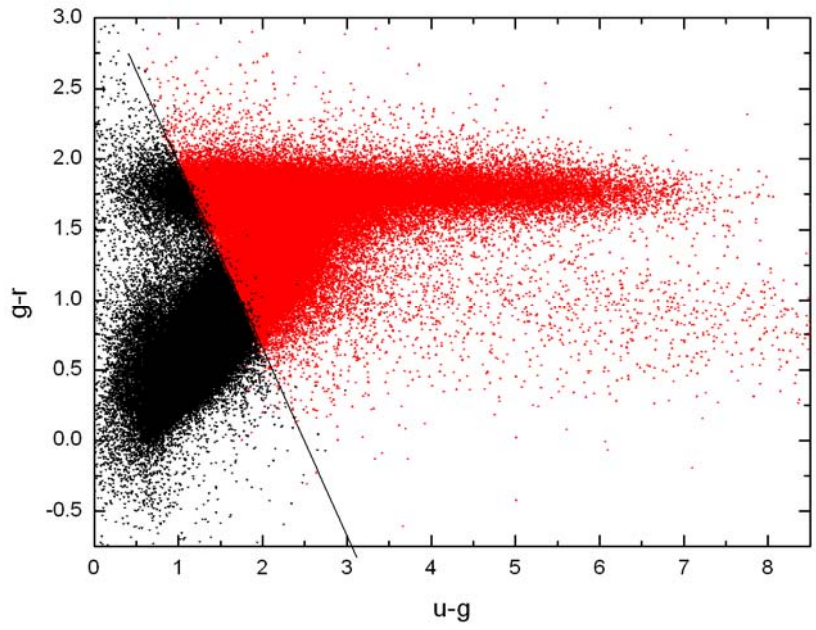

图 5-5　*k-means* 聚类的双色散点图

5.3 测光红移比较

对于传统的模板匹配方法测光谱红移来说，分类和红移问题是耦合在一起的。如果有了天体的分类结果，我们就可以用这类天体的模板进行光谱匹配，求出其红移值；如果知道了天体的红移值，可以将其蓝移到零红移状态，容易进行谱线证认，由证认出的谱线可以知道天体的各种性质。因此知道红移，就可以推算出天体的各种性质和信息，但是如何提高测红移的精度，尤其对测光红移而言，是一个极其有意义的探讨。我们这里提出一个方案，就要研究如何提高天体测光红移精度，其思路是：在对所有星系样本进行哈勃类型粗分类后，对所得的早型星系和晚型星系分别作测光红移预测，计算二者结合后的总样本预测精度，将其与直接对总样本进行红移估测的精度相比较，看是否新的方案能够提高测光红移的精度？

前面我们已经使用 *k-means* 聚类算法将总的星系样本粗分为两类：早型星系和晚型星系样本。这里将从这两类样本数据中选出满足一定标准的做测光红移实验，所用选择标准与上一章相同： *r* 波段 *Petrosian* 星等亮于 17.77 等，光谱红移置信度大于 0.95 并且没有警告标记（即 *zWarning*=0）。这样从原来的 582, 257 个样本中选出 375, 929 个满足要求的，其中包含早型星系 191, 200 个和晚型星系 184,729 个。

利用这两类样本数据，采用神经网络方法来预测测光红移。从早型星系样本中随机抽取 80,000 个用于训练、20,000 个用来评估、余下的 91,200 个作为测试样本。同样晚型星系也选 80,000 个训练、20,000 个评估和 84,729 个作为测试。分别对早型星系和晚型星系做测光红移预测，得出各自的预测精度，最后根据各类的预测结果计算出它们的结合精度，公式如下：

$$\sigma = \sqrt{\frac{1}{N_1 + N_2}\left(\sum_{i=1}^{N_1}(zNN_i - zspec_i)^2 + \sum_{i=1}^{N_2}(zNN_i - zspec_i)^2\right)} \qquad (5\text{-}1)$$

其中N_1=91,200 是早型星系中测试样本的个数，N_2=84,729 为晚型星系中测试样本的个数，*zNN*为神经网络预测红移值，而*zspec*为光谱红移值，两者之间的剩余标准偏差为_。

实验所选的输入参数分别为 SDSS 巡天数据中五个波段的三种不同星等：*Petrosian* 星等、模型（*model*）星等和红化校正（*dereddening*）星等，与另外两个参数 *PetR50* 和 *PetR90* 的结合。网络结构的选取是通过反复尝试，尽量使之最优化。这样根据神经网络给出的早型星系和晚型星系红移的预测值，分别求出两类的测量偏差，计算两者的结合偏差，越小越好。

为了与最通常的方案比较，我们也用神经网络方法直接对总的星系样本进行了红移预测。从总星系样本 375,929 个中随机抽取 150,000 个用于训练，50,000 个样本用于评估，而余下的 175,929 个作为测试样本。使用与上面方案相同的输入参数进行预测红移，得出总样本的直接偏差。对实验中使用的参数、所选网络结构及预测红移的偏差列入下表 5-1。

表 5-1 神经网络方法预测红移实验列表

所选参数	星系类型	所选网络结构	各类偏差	结合偏差	直接使用总样本偏差
Petrosian u,g,r,i,z	早型	5:10:1	0.022105	0.0254	0.026146
	晚型	5:5:5:1	0.028576		
Petrosian u,g,r,i,z, PetR50,PetR90	早型	7:5:10:1	0.020756	0.024	0.024676
	晚型	7:5:10:1	0.027093		
Model u,g,r,i,z	早型	5:10:1	0.01777	0.0204	0.021431
	晚型	5:10:1	0.022968		
Model u,g,r,i,z PetR50,PetR90	早型	7:5:10:1	0.016581	0.0192	0.020282
	晚型	7:5:10:1	0.021738		
dereddening u,g,r,i,z	早型	5:10:1	0.017412	0.0203	0.021094
	晚型	5:10:1	0.022986		
dereddening u,g,r,i,z PetR50,PetR90	早型	7:5:10:1	0.016408	0.0192	0.019553
	晚型	7:5:10:1	0.021739		

通过上表，我们可以明显地看出将早型星系和晚型星系分开

作测光红移预测，所得的结合偏差要比直接使用神经网络对总的星系样本预测红移所得偏差小。可见分别对早型星系和晚型星系作测光红移所得到的预测精度更高。同时在分别对两类型星系作测光红移发现：神经网络对早型星系的预测相当精确，而对晚型星系则要差些。再者，在使用三种不同星等为主要参数时，同上章的结论相似：采用红化校正（*dereddening*）星等作为参数是最好的选择，它对星系红移的预测值比用 *Petrosian* 星等和模型（*model*）星等作为参数的精度都要高，而模型星等的预测结果又优于使用 *Petrosian* 星等的结果。当在各波段五个星等为参数的基础上增加 *PetR50* 和 *PetR90* 两个参数时，对相应的每组实验效果都有所改进，如果用星系总样本直接预测红移，最好的参数组合是使用红化校正星等的七个参数；而将星系分类之后的最好的总精度对应的参数组合为模型星等的七个参数或红化校正星等的七个参数，其σ值均为 0.0192。红化校正星等的七个参数实验测红移的结果如图 5-6 至 5-8 所示，其中 R 表示神经网络输出与目标输出的相关系数，它越接近于 1，表示网络输出与目标输出越接近，网络性能越好。

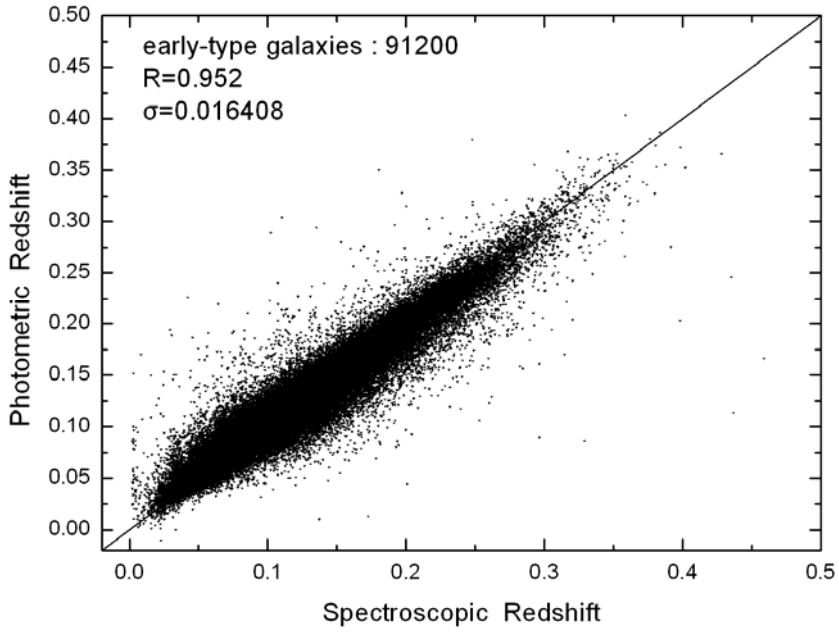

图 5-6　使用红化校正星等七个参数对早型星系测红移

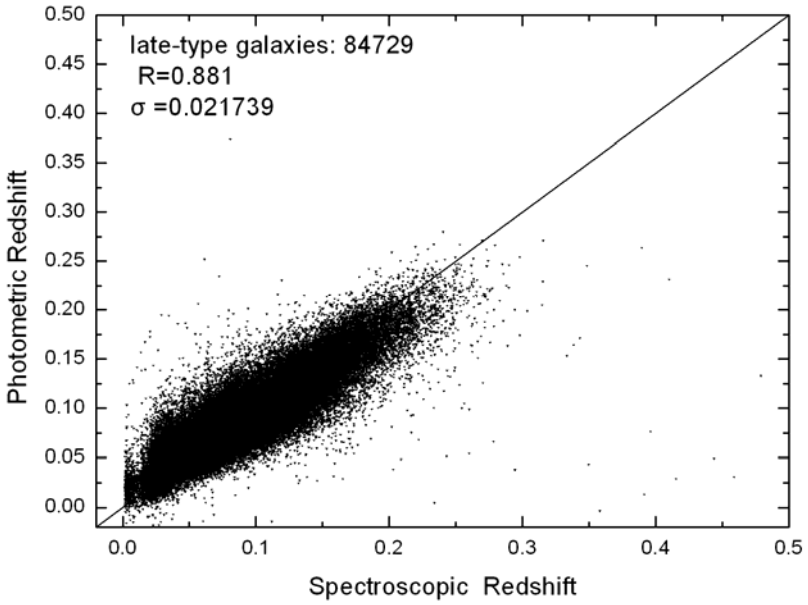

图 5-7　使用红化校正星等七个参数对晚型星系测红移

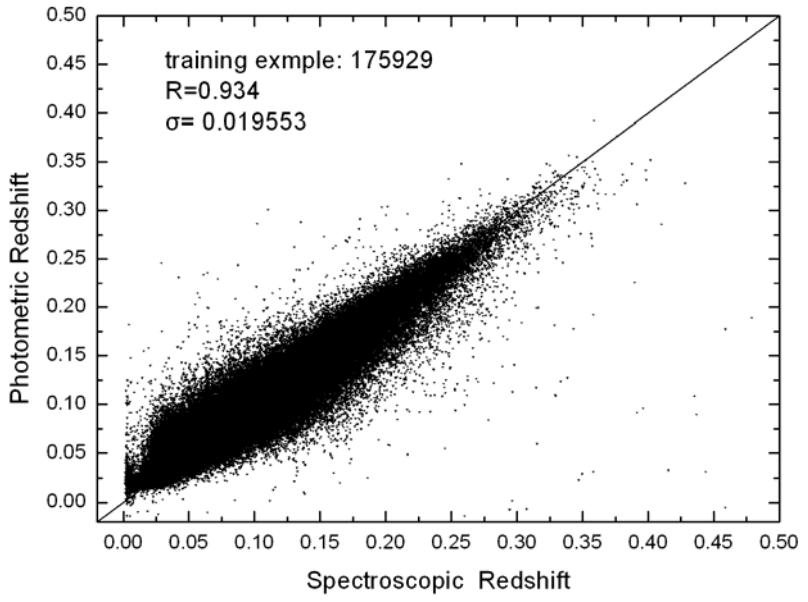

图 5-8　使用红化校正星等七个参数对所有星系测红移

5.4 结论

本章我们主要是先使用 *k-means* 自动聚类算法将 SDSS 巡天数据中的所有星系样本按哈勃类型粗分为两类：早型星系和晚型星系，然后分别对两者预测测光红移。结果表明：当用神经网络预测红移时，发现将早型星系和晚型星系分开后的结合偏差要比直接用总的星系样本来预测红移的偏差小，而且对早型星系的预测精度要远远优于对晚型星系的精度。直接对星系总样本预测红移时，最好的参数组合是使用红化校正星等的七个参数；而对应于星系分类之后的最好总精度的参数组合为模型星等的七个参数或红化校正星等的七个参数，σ值均为 0.0192。因而对测光红移而言，先对样本进行分类可以有效地提高测量精度。

为了与前人的工作比较，这里列出了所有使用神经网络对 SDSS 巡天数据预测红移的结果如表 5-2。从表中可以看出，神经网络作为近几年发展的一种人工智能方法已被应用到了 SDSS 巡天观测释放的各期数据中，尤其在对大样本测光红移的应用方面，神经网络所给出的预测精度是相当可观的，可以作为评价其它测红移方法的参考标准。同时随着研究的深入，使用神经网络预测红移的精度也在不断地提高。2003 年 Firth 等人对 SDSS 早期释放数据（EDR）预测的结果偏差为 0.023[65]；2004 年 Vanzella 等对 DR1 红移的预测偏差也为 0.022[66]；本人前期使用多层感知神经网络对 DR2 预测红移给出的结果偏差为 0.020184[101]；而在 Raffaele 等人的文章中，对两类星系样本（即普通星系和亮红星系）按照一定的红移范围来分成远距星系（z<0.25）和近距星系(0.25<z<0.5)，使用多层神经网络方法分别对远距和近距星系预测红移，给出了普通星系的预测偏差为 0.0208[102]。而我们对 SDSS 第五次释放的全体数据（DR5）采用了非监督的自动聚类算法，让数据按照其自身的特性来聚为早型星系和晚型星系，无人为的干预，完全体现了数据挖掘中的自动化特性。同时使用神经网络方法分别对早型星系和晚型星系预测红移，计算出结合精度，并与直接使用总样本预测红移作比较，发现这种方法提高了测光红移精度，最好的σ值达到 0.0192。因此我们提出的方案是有利于提高测光红移预测精度的，具有一定的意义和应用价值。

表 5-2 使用神经网络预测红移结果比较

参考文献	使用方法	所用数据	预测偏差 σ
Firth *et al.* (2003)	ANNs	EDR	0.0230
Collister& Lahav (2004)	ANNz	EDR	0.0229
Vanzella *et al.* (2004)	MLP	DR1	0.0220
以前的工作	MLP	DR2	0.0202
Raffaele *et al.* (2006)	MLP	DR5-GG	0.0208
目前工作	MLP	DR5	0.0192

第六章 总结与展望

本论文从大规模数字巡天项目的开展谈起，阐述了天文中数据挖掘的必要性与必然性。结合天文学的特点和要求，描述了数据挖掘的一些相关知识、任务及方法。在这些任务中，分类和回归（即预测）是我们要完成的。文中采用了多种数据挖掘方法：*K* 近邻分类法、*k-means* 聚类法与人工神经网络方法，侧重于人工神经网络方法的研究及应用。

第二章详细介绍了人工神经网络的原理及特性，着重阐述了神经网络适用于天文学应用的主要特性及其在天文中的主要应用模型，并列举了它在天文方面的应用实例。说明了神经网络对处理具有复杂特性的天文数据的一些优势，因此它受到国内外天文学家的青睐，已被广泛地应用到各个领域。

第三章是使用神经网络方法来对 SDSS 巡天数据预测测光红移。描述了测光红移相对于光谱红移的优势及传统测光红移所采用的一些方法，简单介绍了 SDSS 巡天数据项目释放的数据及其有关参数。对 SDSS 巡天数据预测测光红移时，主要使用三种有效的星等参数：*Petrosian* 星等、模型（*model*）星等和红化校正（*dereddening*）星等，进行三组实验，并且每组实验又将这些星等参数与星系颜色和各波带处 50%和 90%的 *Petrosian* 流量处半径相结合，通过实验比较得出对神经网络测红移而言最好的一组参数组合。

第四章我们使用数据挖掘中的 *K* 近邻方法来完成对天体的自动分类任务。由天体自动分类的意义引出我们的工作，根据不同天体的物理特性如形态和能谱分布在各波段的差异，在多维参数空间中将活动星系核从恒星和星系样本中分离出来。实验的样本是由多个波段数据库交叉证认获得的，所选参数包含天体在各波段的相关信息。使用 *K* 近邻方法构造分类器，完成自动分类任务，所得分类精度达 97.73%，说明 *K* 近邻方法作为自动化的分类方法是比较合理的、可行的。

最后我们又提出对 SDSS 巡天数据中的星系样本先分类再预测测光红移的方案，将所得测红移的结果与直接使用星系总样本

预测红移的结果相比较，验证我们的方案的可行性与有效性。首先使用 *k-means* 自动聚类算法将所有星系样本按类型分为早型星系和晚型星系，分类结果与已有文献对比，发现这种做法是比较合理的。再使用神经网络方法分别对分类后的样本（即早型星系和晚型星系样本）预测测光红移，计算两者结合后总样本的预测偏差，发现它小于直接使用神经网络对总样本预测红移的偏差，可见我们的方案使神经网络预测测光红移的精度提高了。为天文学家预测测光红移提供了一种新的、比较有效可行的参考方案。

本文我们使用了一些数据挖掘的算法完成了有关的任务，当然数据挖掘的算法远远不止这些，针对不同的问题可以采用不同的算法，但是对同一问题也可以使用多种不同的算法。为了对数据挖掘各种算法深入地研究，在以后的工作中我们将考虑其它方法（如加权回归、支持向量机、径向基函数等等）或者多种方法的结合，实际应用于对天文数据的处理中，完成天体的分类和测光红移任务，以提高分类的准确度和预测红移的精度。

参 考 文 献

[1] 范明、孟小峰译，数据挖掘概念与技术，北京：机械工业出版，2000，14~22

[2] Roberto, T., Giuseppe, L., Milano, L. et al., Neural Networks, 2003, 16~298

[3] 闻新，周露，李翔等，MATLAB 神经网络仿真与应用，北京：科学出版社，2003，285~287

[4] Vanzella, E.，Cristiani, S.，Fontana, A., et al. A&A,2004, 423,761

[5] Ball, N.M.，Loveday, J.,et al., MNRAS, 2004, 348, 1038

[6] Mancini, D., Brescia, M., et al., SPIE, 1997, 3112, 335

[7] Miller, A. S., Coe, M. J., MNRAS, 1996, 279, 293

[8] Maehoenen, P. H., Hakala, P. J., ApJ, 1995, 452,77

[9] Brett, D.R., West, R.G., Wheatley, P. J., MNRAS, 2004, 353,369

[10]Rajaniemi, H. J., Mähönen, P., ApJ, 2002, 566, 202

[11]Hernandez-Pajares, M., Floris, J., MNRAS, 1994, 268, 444

[12]姜玉刚，郭平，计算机科学, 2004(增刊) , 31,7

[13]Bai Ling，Guo Ping，Hu Zhan-Yi, ChJAA, 2005, 5, 203

[14]张彦霞，多波段天体物理中的自动分类方法研究，[博士学位论文], 北京：中国科学院国家天文台，2003，144

[15]钟伟波，金声震，宁书年，计算机工程, 2004, 30, 10

[16]Andreon, S., Gargiulo, G., et al., MNRAS, 2000, 319, 700

[17]Storrie-Lombardi, M. C., Lahav, O. In M.A.Arbib, editor, Handbook of Brain Theory and Neural Networks, 1994

[18]Miller, A. S.，Vistas in Astronomy，1993, 36(2):141

[19]Coryn, Bailer-Jones, [PhD thesis], Cambridge: Institute of Astronomy and Emmanuel College, 1996

[20]Tagliaferri, R., Ciaramella, A., et al. A&A Suppl., 1999,137,391

[21]Bertin, E., Arnouts, S., A&A Suppl., 1996, 117, 393

[22]Jorge, Núñez., Jorge, Llacer., Neural Networks, 2003, 16, 411

[23]Vieira, E. F., Ponz, J. D., ASPC., 1998, 145, 508

[24]Bailer-Jones, C. A. L., Irwin, M., MNRAS, 1998, 298, 361

[25]Snider, S., Allende, P. C., et al. ApJ, 2001, 562, 528

[26]Goderya, S., Lolling, S., Ahmed, R., A&ASuppl., 1999, 31, 832

[27]Naim, A., Lahav, O., Sodre, L. Jr, et al. MNRAS, 1995, 275, 567

[28]Collister, A. A., Lahav, O., PASP, 2004, 116, 345

[29]Borda, F. R. A., Mininni, P. D., Mandrini, C. H., et al., Sol. Phys., 2002, 206(2), 347

[30]Veselovskii, I. S., Dmitriev, A. V., et al, AV, 2000, 34(2), 116

[31]Fuentes, O., Experimental Astronomy, 2001, 12(1), 21

[32]Balastegui, A., Ruiz-Lapuente, P., MNRAS, 2001, 328, 283

[33]Baccigalupi, C., Bedini, L., Burigana, C. et al. MNRAS, 2000, 18, 769

[34]Funaro, M., Oja, E., Valpola, H., Neural Networks, 2003, 16, 469

[35]Zhang, Yanxia, Zhao, Yongheng., PASP, 2003, 115, 1006

[36]Gothoskar, P., Khobragade, S., MNRAS, 1995, 277, 1274

[37]Westerhoff, S., et al., Astroparticle Physics, 1995, 4, 119

[38]Rogers, R. D., Riess, A. G., PASP, 1994, 106, 532

[39]Smaregila, R. et al., Neural Networks in Astronomy, Oxford: Pergamon, 1994, 38, 309

[40]Aussem, A., Murtagh, F., Sarazin, M., Vistas in Astronomy, 1994, 38, 357

[41]Ozard, S., Morbey, C., PASP, 1993, 105, 625

[42]Lloyd-Hart, M., Wizinowich, P., et al., ApJ, 1992, 390, 41

[43]Johnston, M. D., Technical report, Goddard Space Flight Center, NASA, Space Network Control Conference on Resource Allocation Concepts and Approaches, 1991, 183

[44]Faundez-Abans, M., Ormeno, M. I., et al. A&AS., 1996, 116, 395

[45]Baum, W.A., 1962, IAU Sym., 15, 390

[46]Bolzonella, M., Miralles, J.M., Pelló, R., A&A, 2000, 363, 476

[47]Sowards-Emmerd, D., McKay, T. A., Sheldon, E., Smith, J. A., A&A Suppl., 1999, 194, 0410

[48]Yahata, N., Lanzetta, K.M., Chen, H.W., PASP, 2000, 112, 691

[49]Wang, Y. Bahcall, N., Turner, E. L. 1998, AJ, 116, 2081

[50]Corbin, M. R., Vacca, W. D., 2000, AJ, 119, 1062

[51]Fontana, A., D'Odorico, S., 2000, AJ, 120, 2206

[52]Hogg, D. W., Cohen, J. G., Blandford, R., 1998, AJ, 115,1418

[53]Connolly, A. J., Csabai, I., Szalay, A. S., 1995, AJ, 110, 2655

[54]Sowards-Emmerd, D., Smith, J.A., et al. 2000, AJ, 119, 2598
[55]Wadadekar, Yogesh, 2005, PASP, 117, 79
[56]Collister, Adrian A., Lahav, Ofer, 2004, PASP, 116, 345
[57]Firth, Andrew E., Lahav, Ofer, 2003, MNRAS, 339, 1195
[58]Vanzella, E., Cristiani, S., et al, 2004, A&A, 423, 761
[59]York, D. et al., 2000, AJ, 120, 1579
[60]Stoughton, C., et al., 2002, AJ, 123, 3487 (EDR)
[61]SDSS collaboration, 2003, AJ, (DR1)
[62] http://www.sdss.org
[63]Petrosian, V., 1976, ApJ, 209, 1
[64]Tagliaferri, R., Longo, G., et al., 2002, astro-ph/0203445
[65]Firth, A. E., Lahav, O., et al., 2003, MNRAS, 339, 1195
[66]Vanzella, E., Cristiani, S., et al., 2004, A&A, 423, 761
[67]Collister, A. A., Lahav, O., 2004, PASP, 116, 345
[68]Gulati, R. K., Gupta, Ranjan, et al., 1994, ApJ, 426, 340
[69]Bailer-Jones, Coryn, A. L.,Irwin, Mike, 1997, MNRAS, 292, 157
[70]Bellas-Velidis, I., Kontizas, E., 1996, Hellenic Astronomical Society, 411
[71]Bailer-Jones, C. A. L., 1997, PASP, 109, 93
[72]Singh, Harinder P., Gulati, Ravi K., 1998, MNRAS 295, 312
[73]von Hippel, T., et al., 1994, MNRAS, 269, 97
[74]Sodré, L. Jr., Cuevas, H. 1994 , Vistas in Astronomy, 38, 287
[75]Adams, A., Woolley, A., 1994, VA, 38,273
[76]Nielsen, M. L., Odewahn, S. C., 1994, AAS, 26, 1498
[77]Storrie-Lombardi, M. C., Lahav, O. et al. 1992, MNRAS, 259, 8
[78]Odewahn, S. C., Nielsen, M. L., 1994, VA, 38, 281
[79]Philip, N. S., Wadadekar, Y., Kembhavi, A., et al., 2002, A&A, 385, 1119
[80]Odewahn, S. C., Stockwell, E. B., Pennington, R. L. et al.,1992, AJ, 103, 318
[81]Weir, N., Fayyad, U., et al., PASP, 1995, 107, 1243
[82]Weir, N., Fayyad, U., Djorgovski, S. G., 1995, AJ, 109, 2401
[83]Fayyad, U., Smyth, P., Weir, N., et al., Intel. Inf. Sys. 1995, 4, 7
[84]Goebel J, Volk K, Walker H, et al. A&A, 1989, 222, 5

[85]De Carvalho, R., Djorgovski, S. G., et al., ASP Conference Series, 1995, 77, 272

[86]Yoo, J., Gray, A., Roden, J., et al., ASP Conference Series, 1996, 101, 41

[87]Brunner, R. J., Prince, T., et al. 2001, ASPC, 225, 135

[88]Adanti, S., Battinelli, P., et al., 1994, A&A Suppl. 108, 395

[89]Connolly, A. J., Szalay, A. S., et al., 1995, AJ, 110, 1071

[90]Connolly, A. J., Szalay, A. S., 1999, AJ, 117, 2052

[91]边肇祺，张学工等，模式识别，北京：清华大学出版社，2000，140

[92]Zhang, Y., Zhao, Y., 2004, A&A, 422, 1113

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

[94]De Vaucouleurs, G., de Vaucouleurs, A., et al., 1991, Third Reference Catalogue of Bright Galaxies (RC3)

[95]赵刚，陈玉琴译，星系天文学，北京：科学技术出版社，2004，146~151

[96]赵丽芳，[硕士学位论文]，南京：南京师范大学，2003, 29

[97]Nakamura, O., Fukugita, M., et al., 2003, AJ, 125,1682

[98]Shimasaku, K., Fukugita, M., et al., 2001, AJ, 122, 1238

[99]Strateva, I., et al., 2001, AJ, 122, 1861

[100] Blanton, M.R. etal., 2003, ApJ, 592, 819

[101] Li, L., Zhang, Y., Yang, D., et al., 2006, ChJAA, accepted (astro-ph/0612749)

[102] Raffaele, D'Abrusco, Antonino, Staiano, et al., 2007, ApJ, submitted (2007astro.ph/3108R)

致　　谢

在此论文完成之际，回首以往的学习科研生活，很是值得欣慰。整个过程中，我承受了不少的关爱、鼓励和帮助。借此机会，我由衷地表达发自内心的感激之情和最真挚的谢意！

首先，对我的导师杨大卫教授和国家天文台的赵永恒研究员、张彦霞副研究员致以衷心的感谢和崇高的敬意！非常荣幸，我在攻读硕士期间得到了几位老师的悉心教导和热心帮助。杨老师渊博的学识、深邃的物理思想、兢兢业业的工作和严谨求实的研究作风深深感染着我，催我自新，使我进步。赵老师谦虚诚恳的为人、博学多才的能力、宽以待人的态度，深深影响着我。张彦霞老师给予了我殷切的关心、不断的鼓励、耐心的指导和帮助，她脚踏实地、勤奋好学的作风，是我学习的楷模。在今后的人生旅途中，我一定铭记三位老师的谆谆教诲，踏踏实实地做好自己的工作，不辜负他们的厚望。

感谢崔辰州老师对我在天文台学习期间的热情关怀和帮助，罗阿理老师的指导和张昊彤老师的模范作用，使我深深难忘。感谢天文台的梁艳春老师，还有 LAMOST 的苏洪均、石火明、李欣、陈英、袁辉、王刚、门力、冯磊、王淑青、贾磊等各位老师的关心的帮助。

感谢国家天文台 LAMOST 全体成员：吴悦、孙士卫、田海俊、刘超、王丹、高丹、路勇、朱黎楠、薛元、姚嵩、魏娜、罗宇、郑征、尹红星、邹思成等人，一个个在我脑海中活灵活现，在此向大家表示谢意，感谢大家对我的热情帮助和殷切鼓励。

感谢张波老师对我学习上的指导和帮助，他宽容谦虚的仁者风范和严谨的治学态度深深地影响着我。感谢李冀老师对我的热情关怀和帮助，他坦诚的话语常萦绕于胸。感谢河北师范大学物理学院各位老师和领导的关心和帮助。

同时衷心感谢学校里的一些同学和好友：张逢春、曹藏文、陈燕萍、杜云霜、尹少英、孟艳、张红、张玉洁、李静、赵淑霞等各位同学，与他们相识，令人愉快，在此向他们表示谢意。

最后忠心感谢我的父母和家人的支持和理解！

论 文 列 表

1.**李丽丽**，张彦霞，赵永恒，杨大卫，神经网络在天文中的应用，天文学进展，2006 年第 24 卷第 4 期发表

2.**Li Lili(李丽丽)**, Zhang Yanxia(张彦霞), Zhao Yongheng(赵永恒), Yang Dawei(杨大卫), Multi-parameter estimating photometric redshifts with artificial neural networks， Chinese Journal of Astronomy and Astrophysics (ChJAA) (已接收)

3.**李丽丽**，张彦霞，赵永恒，杨大卫，*K* 近邻法对天体的自动分类，中国科学杂志社 （审稿中）