分类号_____ 密级_____

UDC _____ 编号_____

中国科学院研究生院

博士学位论文

海量天文数据融合系统的开发与数据挖掘算法的研究

—工具开发与算法研究

高 丹

指导教师 赵永恒 研究员 中国科学院国家天文台

张彦霞 副研究员 中国科学院国家天文台

申请学位级别 博士 学科专业名称 天文技术与方法

论文提交日期 2008 年 5 月 论文答辩日期 2008 年 5 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 教授

Development of Massive Astronomy Data Federation System and Research of Data Mining Algorithms

—Tool Development and Algorithm Research

Gao Dan

Supervisor:

Prof. Zhao Yong-heng & Asso. Prof. Zhang Yan-xia

National Astronomical Observatories

Chinese Academy of Science

May 2008

Submitted in total fulfillment of the requirements for the degree of Ph.D.

in Astronomical technology and method

摘　要

随着天文学“数据雪崩”和“信息爆炸”时代的到来，为了解决天文数据的复杂性、分布性、海量性和多波段性等一列问题，基于数据库的多波段数据的数据融合和数据挖掘研究逐渐成为天文学研究的热点。我们的工作分为四部分：首先，开发了星表数据的自动入库工具，便于天文学家更好地管理和处理海量数据。然后，研究了海量数据的交叉证认算法。比较了多种索引算法，借鉴 B-tree 索引和 HTM 索引方法的优点，首次提出了 HTM 索引分区与 kd-tree 算法找最近邻结合的方案来快速地交叉证认多波段或不同星表的数据，彻底解决了内存限制和算法复杂度高的问题，并保证了较高的证认精度。还研发了海量数据的交叉证认工具，以及将自动入库、交叉证认等工具整合在一起的海量天文星表融合系统（XMaS_VO），借此系统天文学家可以轻松地进行多波段数据的提取和研究工作。另外，探索了 kd-tree、支持矢量机、随机森林等数据挖掘算法，应用了 SDSS、2MASS、USNO、FIRST 和 ROSAT 等巡天数据，进行了天体的分类、测光红移估计、特征加权和提取等研究。研究结果表明这些方法应用于解决海量天文数据的分类和回归问题时是切实可行的、有效的。最后，将数据挖掘算法研究的成果应用在 LAMOST 选源工作中，以 SDSS 巡天数据为例，用所有已知光谱的点源数据作为训练样本，训练 kd-tree，构造出识别类星体的分类器，从而有效地挑选出适合 LAMOST 望远镜观测条件的类星体候选体。

关键词：多波段数据处理与分析，数据融合，数据挖掘，巡天，星表

Abstract

Development of Massive Astronomy Data Federation System and Research of Data Mining Algorithms

Gao Dan（Astronomical Technology & Method）

Directed by Prof. Zhao Yong-heng & Asso. Prof. Zhang Yan-xia

With the era of data avalanche and information explosion in astronomy coming and a series of problems about complexity, distribution, mass and multiwavelength of astronomical data existing, data federation and data mining based on multiwavelength databases gradually become the hot issue of astronomical research. Our work includes four parts: firstly, the tool of putting data into a database in an automated way is developed, which is beneficial to better manage and handle massive data for astronomers. Secondly, cross-match algorithms are explored. Many kinds of index approaches are compared. Based on the advantages of B-tree and HTM index methods, we firstly put forward a scheme that uses HTM for partitioning and kd-tree for finding nearest neighbors during cross-match process. This scheme realizes the cross-match of multiwavelength data or different catalogs at high speed, and solves the problems of memory limit and high computing complexity, moreover, keeps the high precision of cross-match. We exploited the cross-match tool of dealing with massive data, and the XMaS_VO (XMatch System of Virtual Observatories, short for XMaS_VO) system, which integrated the tool of automated database creation and the cross-match tool, etc. The system helps astronomers to conveniently extract and study multiwavelength data. In addition, the algorithms of kd-tree, support vector machines, random forest are investigated with data from SDSS, USNO, FIRST and ROSAT databases. The object classification, photometric redshift estimation, feature selection and weighting are researched. The experimental results indicates that these approaches are applicable and efficient when used for solving classification and regression of massive data. Lastly, the ideas of data mining are in use of preselecting

source candidates for the LAMOST project. Taking SDSS data for example, we trained kd-tree using the pointed sources with known spectra types and constructed the classifier of separating quasars, therefore selected out the quasar candidates fit for the observation condition of the LAMOST telescope.

Key Words: Methods: Data Analysis – Statistical, Data Integration, Data Mining, Astronomical Databases: Miscellaneous – Catalogues – Surveys

目　录

第 1 章 引言.....1

1.1 数据融合和数据挖掘.....1

1.2 海量天文数据应用的现状.....6

1.3 文章的研究内容.....7

第 2 章 海量天文星表数据的自动入库.....11

2.1 天文星表数据库.....11

2.2 星表自动入库工具中 ReadMe 文件标准.....13

2.2.1 ReadMe 文件标准.....13

2.2.2 FITS 文件入库工具.....14

2.3 星表数据库预处理.....17

2.4 星表自动入库工具的实现.....17

2.4.1 功能及局限.....17

2.4.2 程序流程图.....18

2.5 星表搜索服务.....22

2.6 小结.....24

第 3 章 天文数据融合算法的研究和工具的开发.....25

3.1 交叉证认概述.....25

3.1.1 交叉证认定义.....25

3.1.2 国内外的研究情况.....26

3.1.3 交叉证认的步骤和技术难点分析.....28

3.2 交叉证认算法的研究.....29

3.2.1 基于 B-tree 索引的交叉证认.....30

3.2.2 基于 HTM 索引的交叉证认.....32
3.2.3 基于 HTM 索引分区与 kd-tree 找最近邻算法的交叉证认.....36
3.3 交叉证认工具的开发和应用.....36
3.3.1 功能和服务.....37
3.3.2 服务界面.....38
3.3.3 工具程序流程图.....40
3.3.4 自动入库模块.....44
3.3.5 交叉证认模块.....44
3.4 海量星表数据融合系统（XMaS_VO）的实现.....45
3.4.1 XMaS_VO 的功能及特点.....46
3.4.2 XMaS_VO 逻辑过程.....48
3.4.3 XMaS_VO 功能模块.....49
3.4.4 各功能模块的整合.....51
3.4.5 算法实现流程.....53
3.5 交叉证认在 VO-DAS 中的应用.....55
3.5.1 VO-DAS 简介.....55
3.5.2 VO-DAS 中的交叉证认.....56
3.6 测试结果与应用现状.....58
3.6.1 模拟数据检验交叉证认准确率.....58
3.6.2 XMaS_VO 的应用.....61
3.7 小结.....62

第 4 章 天文数据挖掘研究.....63

4.1 分类和回归研究.....64
4.1.1 分类研究.....64
4.1.2 回归研究.....66
4.2 算法原理.....67

4.2.1 kd-tree 算法 67
4.2.2 SVMs 算法 69
4.2.3 随机森林算法 69
4.3 用 kd-tree 和 SVMs 算法分类类星体和恒星 70
4.3.1 巡天 70
4.3.2 样本 71
4.3.3 算法性能评估量 74
4.3.4 结果与讨论 76
4.4 用随机森林算法分类类星体、恒星和星系 83
4.4.1 巡天 83
4.4.2 样本与参数 84
4.4.3 结果与讨论 88
4.5 用 kd-tree 算法评估星系的测光红移 90
4.5.1 样本 90
4.5.2 结果与讨论 90
4.6 小结 92

第 5 章 数据挖掘应用——为 LAMOST 预选类星体候选体 95

5.1 LAMOST 项目简介 95
5.2 各种预选源方法及其成功率 96
5.3 基于 SDSS 巡天数据的 LAMOST 预选源工作 99
5.4 小结 110

第 6 章 总结和展望 111

附录 A　XMaS_VO 各功能模块整合脚本文件 115

附录 B　XMaS_VO 使用详细说明 119

附录 C　XMaS_VO 移植方法 127

附录 D　ExecPlan 详细设计 129

参 考 文 献 141

发表文章 I 147

发表文章 II 149

致 谢 151

表　格

表 3-1　HTM 不同划分级次对应的单元数和单元大小 36
表 3-2　XMaS_VO 应用实例 62
表 4-1　星表样本数量 72
表 4-2　样本参数的统计特性 73
表 4-3　含混矩阵 75
表 4-4　使用 kd-tree 算法 n=11 时不同输入参数准确率的对比 77
表 4-5　使用 kd-tree 算法当输入参数为 u-g、g-r、r-i、i-z、r 时不同 n 值的结果的比较 78
表 4-6　使用 RBF SVMs 算法当输入参数为 u-g、g-r、r-i、i-z、r 时不同可调参数的结果的比较 79
表 4-7　使用 RBF SVMs 算法当参数 c=5、γ=5 时不同输入参数的结果的比较 80
表 4-8　使用 10 倍交叉确认方法当输入参数为 u-g、g-r、r-i、i-z、r 时不同算法不同可调参数的结果的比较 81
表 4-9　样本描述 84
表 4-10　样本 1 参数统计特性 85
表 4-11　样本 2 参数统计特性 86
表 4-12　样本 1 的参数权值 87
表 4-13　样本 2 的参数权值 87
表 4-14　用样本 1 分类类星体和恒星的结果 88
表 4-15　用样本 2 分类类星体和恒星的结果 88
表 4-16　用样本 1 分类类星体、恒星和星系的结果 89
表 4-17　用样本 2 分类类星体、恒星和星系的结果 89
表 4-18　用 kd-tree 算法评估星系测光红移 91
表 4-19　不同测光红移方法的对比 91

表 5-1 SDSS DR6 图像数据 100
表 5-2 SDSS DR6 光谱数据 101

插　图

图 1-1　数据挖掘流程图 3
图 1-2　工作框架图 8
图 2-1　FITS 文件入库工具流程图 16
图 2-2　星表自动入库工具程序流程图 20
图 2-3　星表自动入库工具的入库选择界面 21
图 2-4　录入的数据库示例 22
图 2-5　星表搜索服务界面 23
图 2-6　星表搜索服务示例 24
图 3-1　多波段数据分析流程图 26
图 3-2　B-tree 索引 31
图 3-3　基于 B-tree 索引的交叉证认（1） 32
图 3-4　基于 B-tree 索引的交叉证认（2） 32
图 3-5　HTM 算法天区划分 34
图 3-6　HTM 算法编码方案（1） 34
图 3-7　HTM 算法编码方案（2） 35
图 3-8　基于 HTM 索引的交叉证认 35
图 3-9　交叉证认工具主页面 39
图 3-10　选择参数和交叉证认结果页面 40
图 3-11　交叉证认工具程序流程图 41
图 3-12　一对一的结果 42
图 3-13　VOPlot 可视化结果 43
图 3-14　用户上传自己星表的上载文件页面 43
图 3-15　海量星表融合系统的设计 47
图 3-16　海量星表融合系统的逻辑过程 49
图 3-17　带有交叉证认功能的查询流程图 57

图 3-18　模拟数据噪音 60
图 4-1　三维 kd-tree 68
图 4-2　类星体和恒星抽样样本散点图，前 4 个是不同颜色组合的双色图，后 2 个是前三个主分量空间的分布图 74
图 4-3　公式（1, 2, 3） 75
图 4-4　公式（4, 5, 6, 7） 76
图 4-5　类星体和错分类星体的红移分布 82
图 4-6　类星体和恒星及错分的类星体和恒星的模型 r 星等分布 82
图 5-1　SDSS 巡天中的所有点源随模型 r 星等的统计分布 102
图 5-2　SDSS 巡天中没有光谱观测的点源随模型 r 星等的统计分布 102
图 5-3　SDSS 巡天中具有光谱观测的点源随模型 r 星等的统计分布 103
图 5-4　用分类器预测 QsoTarget 数据表 103
图 5-5　用分类器预测无光谱点源星表 104
图 5-6　QsoTarget 的全部数据和被分类为恒星的数据 i 星等分布比较 106
图 5-7　SDSS 无光谱点源数据基于 i 星等分布（1） 106
图 5-8　SDSS 无光谱点源数据基于 i 星等分布（2） 107
图 5-9　满足 LAMOST 观测条件的类星体候选体的 i 星等分布（19.5<i<21） 108
图 5-10 满足 LAMOST 观测条件的类星体候选体的 i 星等分布（19.5<i<20.5） 108
图 5-11　满足 LAMOST 观测条件的类星体候选体的 i 星等分布（19.1<i<21） 109
图 5-12 满足 LAMOST 观测条件的类星体候选体的 i 星等分布（19.1<i<20.5） 109

图 D 1　ExecPlan 执行计划的流程图（1）.....129
图 D 2　ExecPlan 执行计划的流程图（2）.....130
图 D 3　ExecPlan 执行计划的流程图（3）.....131
图 D 4　ExecPlan 执行计划 From 语句序列图.....132
图 D 5　ExecPlan 执行计划 Select 语句序列图.....133
图 D 6　ExecPlan 执行计划 Where 语句序列图.....134
图 D 7　ExecPlan 执行计划 GroupBy、Having 语句序列图.....135
图 D 8　ExecPlan 执行计划 Having 语句序列图.....136
图 D 9　ExecPlan 执行计划 OrderBy 语句序列图.....137
图 D 10　ExecPlan 执行计划参数名替换序列图.....138
图 D 11　ExecPlan 执行计划生成 ActivityRequest 序列图.....139

第 1 章 引言

随着科学技术的发展，天文学已经步入了数据爆炸、信息丰富的新时代。多波段数据急剧增长，从伽玛射线、X 射线、紫外、光学、红外到射电波段都有观测项目在进行，天文数据的数据量非常大，总量已经达到 PB 量级。通常，新的观测数据能带来新的现象或规律的发现。由于天文数据来自于不同的观测仪器、观测地点和观测时间，而且世界各国的天文资料保存和管理方式存在巨大差异，不同历史时期数据保存和管理方式也存在巨大差别，致使天文数据的类型、格式和结构各不相同，这些都导致了天文数据的复杂性。尽管世界各地的天文数据是可以自由共享的，但是数据结构的混乱和数据存储方式的多样性就像语言不通一样严重地影响着天文学家对这些数据资源的理解和应用，从而天文数据未能得以有效地使用。在这种形势下，天文学家必须开发新的数据存储、管理、分析和挖掘的工具以应对形势发展的需要。针对天文数据的分布性、异源性、多波段等特性，可以开发数据融合工具，以便有效地对多波段数据进行处理和分析；面对天文数据的海量性，需要发展和应用有效的数据挖掘技术来实现对其分析和挖掘。

1.1 数据融合和数据挖掘

数据融合的概念虽始于 70 年代初期，但真正的技术进步和发展乃是 80 年的事，尤其是近几年来引起了世界范围内的普遍关注。虽然数据融合的概念很容易理解，但在各种不同的学科中，它们的精确含义却各不相同：如汇合、组合、协合、综合等，这些概念或多或少有相同之处，然而又各有差异。美国国防部认为：“数据融合是一个多级多侧面的加工过程，包括对多个源的数据和信息的自动化探测、互联、相关、估计和组合处理。”L. A. Klein 又将这一定义提炼为：“对单源或多源数据和信息进行自动探测、互联、相关、估计和组合的多级、多侧面

的加工处理过程。”与数据融合是将各类信息相组合的定义相比，这种定义（多级处理）更具有一般性。天文中的数据融合技术是将多个独立的星表或图象、数据库或数据集通过位置或星名等信息整合成一个整体。例如：图像融合是用特定的算法将两幅或多幅图像综合成一幅新的图像。

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。

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：

（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。

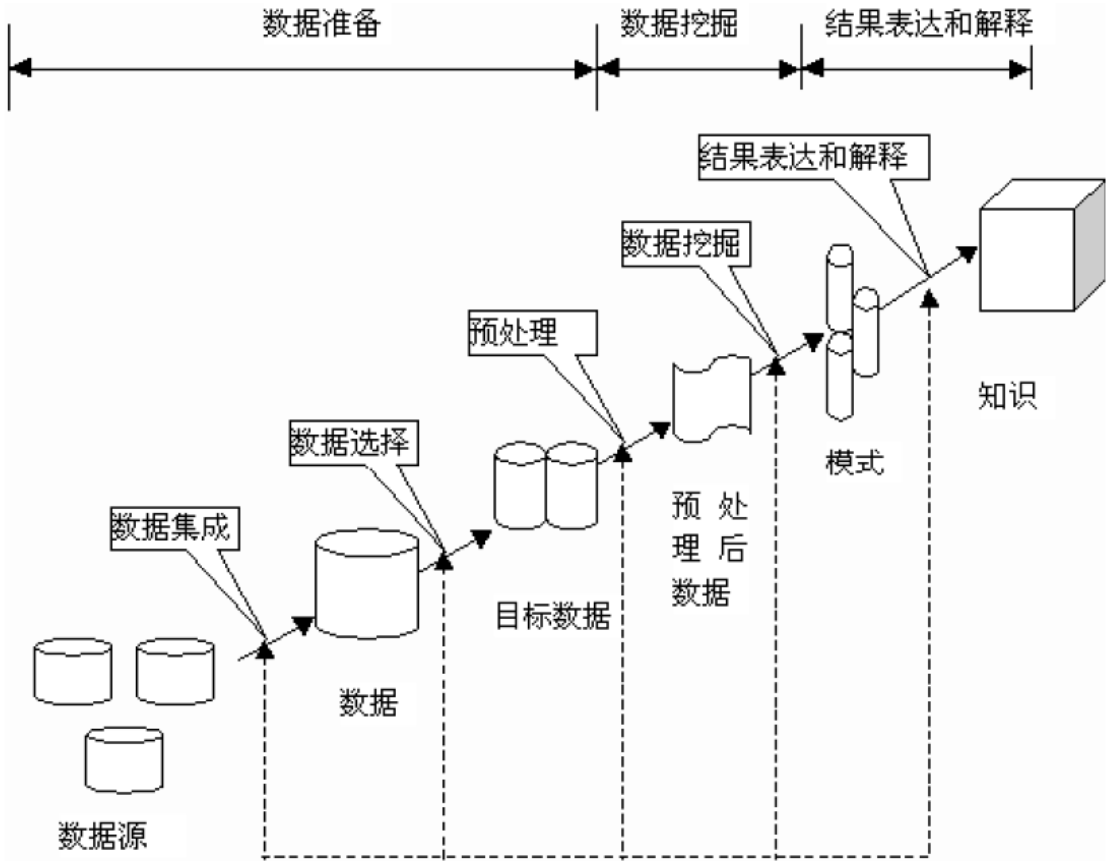


图 1-1 数据挖掘流程图

（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。

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，使用其它方法则可能奏效。

1.2 海量天文数据应用的现状

面对复杂的海量天文数据，如何对其进行分析、处理和管理是全球天文学界面临的重大课题。该课题的解决依赖于信息技术、计算机技术和数据挖掘等技术的发展和应用。相应的数据融合、数据分析、数据挖掘和数据可视化等技术在天文学中的应用逐渐成为海量天文数据应用中的研究热点。

在海量天文数据的处理方面，国外的天文学家开发了很多实用的工具。

在数据融合方面，法国斯特拉斯堡数据中心（CDS）[2] 开发了数据融合工具 VizieR[3] 和数据整合工具 Aladin[4] ，VizieR 实现了对目前已发表数据的多种查询方式，而 Aladin 着眼于将数据可视化。交叉证认方面，比较知名的有 OpenSkyQuery[5] ，用户可以通过 ADQL[6] 查询语言来进行星表的交叉证认，但系统限制查询的行数在 5000 行之内。

数据分析和可视化方面。基于 Java 的 Mirage[7] 可以用表格、直方图、树形结构、图表等多种形式显示坐标点、某类数据点以及相似结构的数据点的投影图，并提供了图表操作来研究同类天体或事件的不同属性之间的相关性以及不同层次的分析之间的纵向联系。另外，在虚拟天文台[8] [9] 框架下开发的 VOStat[10] 提供了面向天文学的统计原型工具，有主分量分析、存在分析、聚类、离群数据探测等功能。值得一提的还有印度虚拟天文台开发的 VOPlot[11] ，它可以对任何星表作图，用户也可对原始数据进行加工，例如增加列和进行一些运算生成新的列等。

但是上述工具都无法解决海量数据的交叉证认和数据融合问题，而海量多波段数据的数据挖掘研究又是以数据融合为基础的。所以，要进行海量多波段数据挖掘研究的首要任务必须实现海量数据融合的工具。

在海量天文数据处理技术发展的过程中，随着数据量的增长，我们需要更好的数据管理和更方便的数据获取及处理服务。天文学家一度是使用文件系统直接对文件进行编程和处理来对海量天文数据进行研究的。随着数据库技术的发展，越来越多的数据库技术被引入到天文学研究中。Jim Gray[12] 对数据库格式和文

本格式做了比较，很明显数据库在性能和便利上都优于文本。

天文中使用数据融合技术是将多个独立的星表或图像、数据库或数据集通过位置或星名等信息整合成一个整体。因为各波段数据是高度相关的，需要在高维参数空间内进行研究。将现有的各波段数据集融合起来，将涌现出全新的、无法预见的、意义重大的科学产出，这是一种仅靠单独使用其中某一部分数据所不能产生的新科学。星表融合的一项重要工作就是基于不同星表内的天体在位置上的相关性，进行星表间的交叉证认工作。

数据挖掘是一种在大型数据库中寻找感兴趣或是有价值信息的过程。具体到天文学上就是从天文数据中提取信息和发现知识。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的天文现象，或者根据数据来区分不同类型的天体等，都需要充分运用在信息科学中迅速发展的数据挖掘和知识发现技术[13] 。关于数据挖掘在天文中的应用可参考文献[13] 。新的大型巡天项目将产生巨量的科学数据。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。运用数据挖掘技术可以有效地解决天文学中的“数据雪崩”问题，这对天文学发展是至关重要的。

1.3 文章的研究内容

本论文的核心是海量天文数据的处理和挖掘，其中主要是天文数据融合系统的开发和数据挖掘算法的应用。天文数据具有格式不统一、海量、分布于世界各大天文数据中心等特点。要将海量的、复杂的、多波段的、异构的天文数据有效地融合起来，需要自动化的星表入库工具和交叉证认工具；为了提高交叉证认的效率，需要研究海量数据交叉证认的算法；要从交叉证认的多维数据中提取信息和知识，需要合适的数据挖掘算法。

图 1-2描述了本论文的整个工作框架。

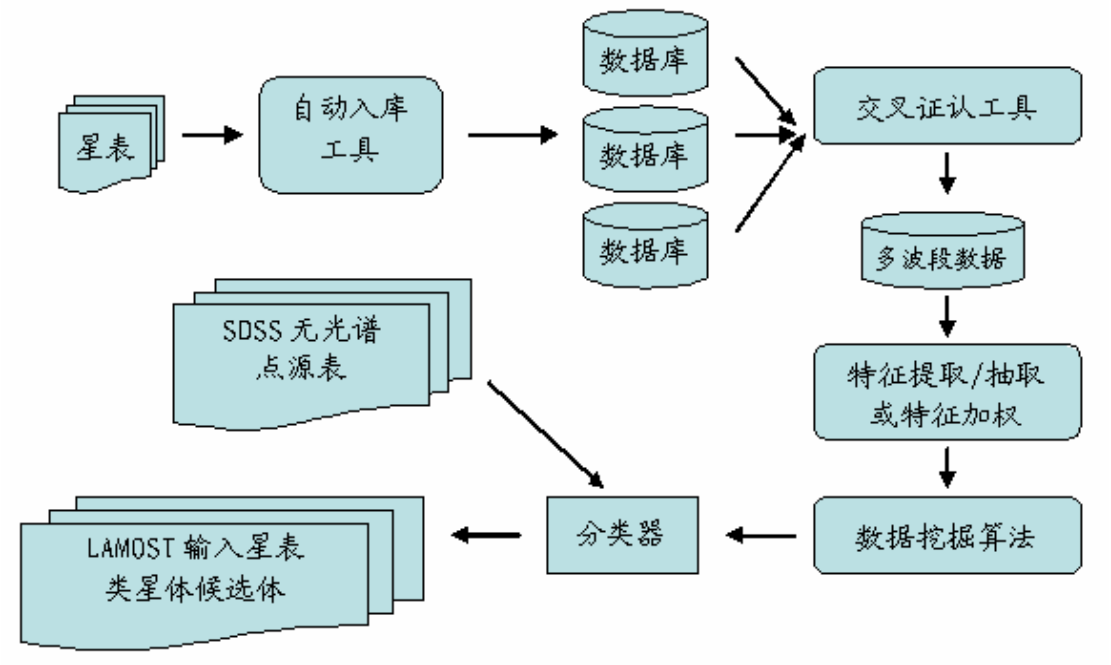


图 1-2　工作框架图

为了更好地管理和处理海量数据，解决天文数据的复杂性和分布性的问题，我们实现了基于星表 ReadMe 文件的星表自动入库工具，用户只需根据模板创建 ReadMe 文件，即可自动入库。自动入库工具还将星表数据统一进行预处理，统一赤经赤纬格式，使数据融合工具可以对处理过的星表数据进行统一的管理和操作。

为了实现海量数据的交叉证认，提高交叉证认的效率，我们对数据库尝试了 B-tree 索引和 HTM（Hierarchical Triangular Mesh，简称 HTM）索引，并借鉴这两种索引方法的优点，引入了 kd-tree（k-dimensional tree）算法找最近邻的思想，创新地提出了 HTM 索引分区与 kd-tree 找最近邻算法结合的方案，彻底解决了内存限制和算法复杂度高的问题，并保证了高的证认精度。借鉴国外开发的交叉证认工具的优点，我们开发了交叉证认工具，实现了海量数据的交叉证认。为了将自动入库、交叉证认等工具整合起来批处理地、多任务地进行工作，我们开发了海量天文星表融合系统 XMaS_VO（Xmatch System of Virtual Observatory，间称

XMaS_VO），天文学家可以通过该融合系统轻松地进行多波段数据的提取和研究。使用模拟数据和实测数据的测试证明了 HTM 索引分区与 kd-tree 找最近邻算法的有效性和 XMaS_VO 系统的可靠性、高效性、灵活性目前已经有许多基于 XMaS_VO 系统的海量星表的交叉证认应用。

XMaS_VO 系统使海量天文数据挖掘课题研究成为可能，并使占数据挖掘工作量 60%的数据准备和预处理工作变得非常方便，算法研究人员只需关注算法性能和应用研究本身，而无需对海量数据处理投入太多精力。我们探索了几种数据挖掘算法（kd-tree、支持矢量机、随机森林等），基于 SDSS（the Sloan Digital Sky Survey）、2MASS （the Two Micron All-Sky Survey)、USNO、FIRST 和 ROSAT 等巡天的多波段数据，进行了天体分类、测光红移估计、特征加权和提取等研究。研究了数据从选择、清洗、降维去噪、特征选择和特征提取、按照挖掘任务选取挖掘方法到结果的解释与评估，并从多波段数据中利用数据挖掘的方法为 LAMOST 挑选出类星体候选体，或一些特殊的源。

算法研究实验中类星体和恒星的分类效率均高达 90%以上，因而可以认为这些方法应用于分类问题是切实可行的、有效的。因此，结合 LAMOST[14] 的观测条件，将数据挖掘算法研究的成果应用在 LAMOST 选源工作中，以 SDSS 巡天数据为例，用所有已知光谱的点源数据作为训练样本，训练 kd-tree，构造出识别类星体的分类器，成功地挑选出适合 LAMOST 望远镜观测条件类星体候选体。这些方法的成功应用将有助于天文学家处理和分析数据，发现知识，以此来带动天文理论和技术的发展和完善，而且这对大型巡天项目 LAMOST 的输入星表和光谱处理有着非常重要的实用价值。

第 2 章　海量天文星表数据的自动入库

天文数据主要包括星表、图像、光谱、文献资料等，其中星表是包含天体信息的数据表格，是天文学家最常用的天文数据。大部分天文数据都经过良好组织和归档，当前世界上有许多天文数据中心在数据的整理、归档和发布等方面做了大量的工作，并提供其数据集合的基本查询服务，用户可以通过互联网进行访问获得需要的数据。中国最大的天文数据中心是北京天文数据中心（Beijing Astronomical Data Center，简称 BADC）[15] ，其数据库的重要部分是天文星表数据库。

目前天文数据分布地存储在全球上千地点的数据中心，访问和管理这些高度分散、复杂、海量的天文数据是非常麻烦的。随着天文学“数据雪崩”时代的到来，为了实现全球天文数据资源的融合，使天文学家能够高效快捷的获取数据、组织数据、分析数据，虚拟天文台[8] [9] 便应运而生。虚拟天文台是一个通过先进的信息技术把全球范围内的研究资源无缝透明联结起来组成的数据密集型在线天文研究环境。它实现分布式数据库的统一访问机制，并提供高效处理海量数据的处理软件、分析工具和可视化工具等。然而虚拟天文台应用效果的好坏直接依赖于它的底层建筑——天文星表数据库。

2.1　天文星表数据库

天文数据库是天文学家观测和研究成果的结晶，是现在人们进行天文观测和研究的综合型数据库。数据来源于全世界发表的星表数据、图像数据和天体测量数据等。天文数据库不但供天文观测和天文研究使用，而且在大地测量、卫星跟踪和遥控及天文学和地球科学的交叉研究中，也是不可缺少的参考数据。

星表数据的大小差别悬殊，小到几十个目标，大到上亿甚至上十亿个目标，

同时近年来随着巡天项目增加和扩大而得到的大星表越来越多，为了实现对这些星表的高效管理和查询，中国虚拟天文台[16] [17] 设计实现一个功能完善的天文星表数据库管理系统。天文星表数据库的主要数据来源于斯特拉斯堡天文数据中心(CDS)负责收集整理的天文星表，这些星表包括银河系内和银河系外恒星、星系、活动星系核以及其他类型的天体信息，也收录了太阳及太阳系内天体的观测数据、测量数据和物理数据。CDS 目前收录的 6000 多个星表通过 ASCII 码文本格式或 FITS(Flexible Image Transport System)[18] 格式文件发布，这些文件可以通过 FTP 下载。

天文星表数据库包括“数据”库和“元数据”库两部分，“数据”库是由导入的星表数据组成，也称星表数据库，“元数据”库管理这些星表的有关元数据信息，也称星表元数据库。星表都是经过观测、处理、整理、归档的数据，由于观测项目、观测目标、观测仪器、观测波段、观测时间等的不同，其格式也非常不统一。元数据是描述数据的数据，用来描述档案、档案提供的服务、其中的数据集合、每个数据集合的结构和语义（即含义）、以及数据集合中每个数据集的结构和语义。从数据中心下载下来的星表是一个文件夹，包括 ReadMe 文件（即星表元数据）、数据表文件和其他文件，都是文本格式。其中 ReadMe 文件以标准格式提供星表的有关信息，如包含哪些文件，并提供数据表文件中每一列的位置、名称、单位等描述（即表结构）。一个星表对应着一个数据库(即 database)，一个星表可能包含多个数据表。星表的一个数据表对应着数据库的一个表(即 table)。

以前，星表数据库的入库工作只实现了半自动入库，即只能根据星表 ReadMe 文件人工修改入库程序；星表元数据信息收集整理工作也主要由数据库录入员手工进行，并按照国际虚拟天文台联盟(IVOA)[19] 制定的资源元数据标准设计了星表元数据库[20] 。为了便于天文学家方便快捷地创建自己的星表数据库，有效地管理数据和应用数据，我们研究开发了星表自动入库工具。同时我们通过从星表 ReadMe 文件中自动提取星表的元数据信息录入到星表元数据库中，并提供星表搜索服务[21] ，让天文学家方便地根据自己的需求查找到相应的星表数据。

2.2 星表自动入库工具中 ReadMe 文件标准

星表自动入库工具实现了自动读取星表 ReadMe 文件并将表自动录入数据库的功能，完全屏蔽掉了星表 ReadMe 文件及数据表本身的细节，提供友好的用户界面让用户快速、方便地入库。如果用户自己处理了一批数据，也可以根据我们制定的 ReadMe 文件标准编写简单的 ReadMe 文件。使用星表自动入库工具的用户主要有以下三种：天文数据的维护者、基于大型数据集的研究工作者和一些需要基于星表数据库进行数据开发的程序员。下面首先介绍星表自动入库工具中的 ReadMe 文件标准和数据文件格式问题。

2.2.1 ReadMe 文件标准

ReadMe 文件必须符合标准要求才能用于自动入库。如不符合，比如列名有特殊字符或有重复的情况，就需要手动修改 ReadMe 文件；需要修改单位的情况，比如赤纬角秒单位是 0.1S，应改为 S，修改 ReadMe 文件的同时，相应的数据文件也需要修改。

ReadMe 文件标准：

1) ReadMe 文件必须是文本文件。

2) 文件中必须含有描述星表文件的“File Summary:”部分和描述星表列信息的“Byte-by-byte ”（也可能为“Byte-per-byte ”）部分。其中“File Summary:”描述了星表文件的信息，包括名字、行长度最大值、行数和文件标题；“Byte-by-byte ”以表格形式描述了数据文件的结构，每一行描述了数据的一列属性信息。

3) “Byte-by-byte ”部分的数据文件名（如 *.dat）必须和数据文件对应。

4) 赤经赤纬列的 Label 必须为 RAdeg、DEdeg（度）或 Rah、Ram、RAs、DE-、DEd、Dem、DEs（时分秒），时分秒形式精度可以只到分（即 Rah、Ram、DE-、DEd、DEm）。

5) 如果星表中有两个历元，J2000 历元必须在 B1950 前面。必须有 2000 年的历元。如只有 1950 年的历元，可调用历元转换程序转换成 2000 年的历元，

然后将转换后的坐标数据改入数据文件，再修改 ReadMe 文件的列描述信息并入库。

6) 列名 Label 只能包含数字、字母和下划线，列名不能重复。

最简易 ReadMe 文件举例：

```
File Summary:
--------------------------------------------------------------------------
-----------
 FileName    Lrecl    Records    Explanations
--------------------------------------------------------------------------
-----------
ReadMe         80          .    This file
first.dat     124     811117    North & South Galactic Caps
--------------------------------------------------------------------------
-----------

Byte-by-byte Description of file: *.dat
--------------------------------------------------------------------------
-----------
   Bytes Format Units   Label     Explanations
--------------------------------------------------------------------------
-----------
   1- 16  A16   ---     FIRST     *FIRST Source designation
  18- 19  I2    h       RAh       *Right Ascension J2000 (hours)
  21- 22  I2    min     RAm       *Right Ascension J2000 (minutes)
  24- 29  F6.3  s       RAs       *Right Ascension J2000 (seconds)
     31  A1    ---     DE-       *Declination J2000 (sign)
  32- 33  I2    deg     DEd       *Declination J2000 (degrees)
  35- 36  I2    arcmin  DEm       *Declination J2000 (minutes)
  38- 42  F5.2  arcsec  DEs       *Declination J2000 (seconds)
     44  A1    ---     Wflag     *[W] Warning flag
  46- 53  F8.2  mJy     Fpeak     *Peak flux density at 1.4GHz
  56- 63  F8.2  mJy     Fint      *Integrated flux density at 1.4GHz
  66- 71  F6.3  mJy     Rms       *Local noise estimate
  73- 78  F6.2  arcsec  MajAxis   *Major axis (FWHM)
  80- 85  F6.2  arcsec  MinAxis   *Minor axis (FWHM)
  87- 91  F5.1  deg     PA        *Position angle
  93- 98  F6.2  arcsec  fMaj      *Fitted MajAxis before deconvolution
 100-105  F6.2  arcsec  fMin      *Fitted MinAxis before deconvolution
 107-111  F5.1  deg     fPA       *Fitted PA before deconvolution
 113-124  A12   ---     Field     *Name of the coadded image containing
                                    the source
--------------------------------------------------------------------------
-----------
```

2.2.2 FITS 文件入库工具

星表数据入库工具的数据文件格式必须是 ASCII 文件。如果是其他文件格

式，比如 FITS，则调用程序将 FITS 文件转换成 ASCII 文件，然后再入库。其他文件格式也要先转换成 ASCII。为了实现这个功能，我们开发了 FITS 文件入库工具。

FITS (Flexible Image Transport System)是天文学界常用的数据格式，它专门为在不同平台之间交换数据而设计的。1982 年，国际天文联合会 (IAU) 确定 FITS 为世界各天文台之间用于数据传输、交换的统一标准格式。

FITS 描述了数据定义和数据编码的一般方法，被用于多维矩阵数据（比如一维光谱、二维图像、三维数据立方体）、表列数据的传输、分析和存档。它是与机器无关的，提供了图像的单值转换，精度包括符号在内可以达到 32 位。

当FITS中的数据是星表的时候，要将此FITS文件包含的星表数据进行处理，常常需要首先将此数据导入数据库中。因为我们已经开发了 ASCII 文件自动入库的工具。所以本工具只需要将 FITS 文件中的星表数据提取、转换为 ASCII 星表文件即可。

另一方面，当我们只关心此星表的某些列，本工具能够按照用户指定列，只提取这些列，然后连接 ASCII 自动入库的程序，即可只将用户感兴趣的数据导入数据库。

必须指出的是，此工具的特点还在于能够处理非常大的 FITS 文件，特别是将大于 500M 的 FITS 文件入库。因为对于体积庞大的 FITS 星表文件，其星表的行数往往在 10 万量级甚至百万量级。使用现有的办法，就会把特定列全部读取到内存，会导致内存不足的问题。

入库大型 FITS 文件的思路就是将大型文件切割成多个小文件，然后分别读取，再将读取后的内容加以合并。这里的切割是按照行来进行切割。

切割功能由此工具调用 fitscopy 来实现，fitscopy 是由 CFITSIO[22] 提供的一个 Linux 下的命令行程序。CFITSIO 是 C 语言的 FITS 文件输入输出（IO）工具集。

本工具首先根据 FITS 文件的行数，自动判断要将原 FITS 文件切割成多少个小文件，并对这些小文件的文件名进行控制。然后根据用户提供的参数，在各个

小文件抽取用户需要的列，每列分别保存为一个临时文件。最后，将这些列文件逐个逐行读取，合并为一个文件，就是用户需要的 ASCII 格式的 FITS 星表文件。

本工具被分成以下模块：Namespace、FITScutter、FITSRead、ReadCol、FileMerge 等。其中 Namespace 主要是对文件名进行控制；FITScutter 根据文件大小生成批处理程序，该程序通过调用 fitscopy 对文件进行切割；FITSRead 和 ReadCol 进行 FITS 列的抽取；FileMerge 则将小文件进行合并。

本工具在 Linux 上运行。程序运行流程如图 2-1所示。

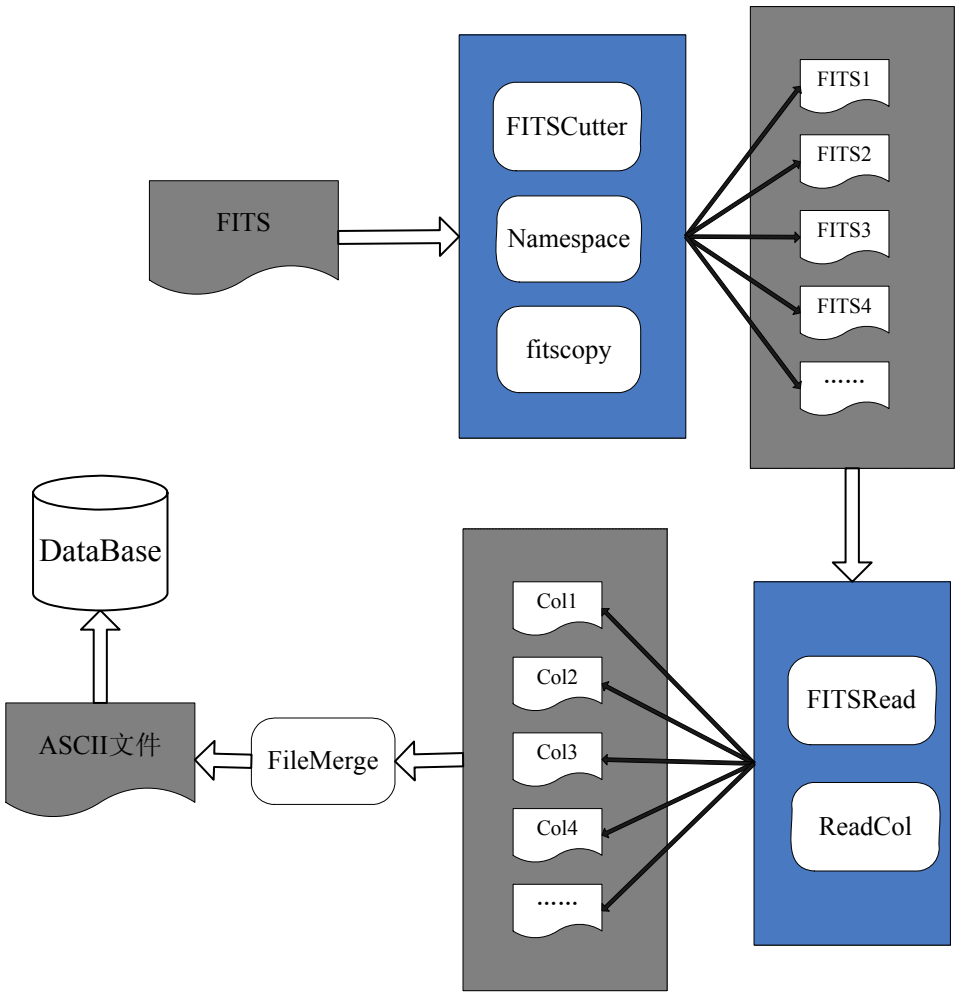


图 2-1　FITS 文件入库工具流程图

2.3 星表数据库预处理

为了进一步的数据融合和数据挖掘，还需将星表数据统一进行预处理，使数据融合工具可以对处理过的星表数据进行统一操作。

我们的星表数据采用数据库的形式统一管理。星表均由自动入库工具根据星表 ReadMe 文件自动入库，并自动进行数据库星表预处理，而交叉证认亦建立在统一管理的数据库星表之上。星表数据库预处理工作整合在自动入库工具中自动进行。

以下是数据库星表预处理的内容：

(1) 统一星表赤经、赤纬标识为度的形式：RAdeg、DEdeg，以便交叉证认工具识别和计算。如果星表赤经、赤纬是度的形式，则统一标识为 RAdeg、DEdeg；如果星表赤经、赤纬是时分秒、度角分的形式，则统一标识为 RAJ2000、DEJ2000 或 RAB1950、DEB1950，并转换成度的形式。在表中加入统一标识为 RAdeg、DEdeg 的两列。

(2) 按 DEdeg 排序，如果用赤经还需要考虑 0 度、360 度的特殊情况，若是赤纬就不用考虑。排序可以提高查询效率。

(3) 加一列主键 id，该主键是逐一递增的整型数据，用以标识每一行数据（即每一个源），这样，为了节约内存，可以只使用主键 id、赤经、赤纬来进行交叉证认和数据融合。在得到的结果中，再通过主键 id 提取其他需要的参数。

(4) 建立多值索引（DEdeg、RAdeg），提高查询效率。

(5) DEdeg、RAdeg 均为非空（not null），因为数据库中索引为 not null 有利于提高性能。

2.4 星表自动入库工具的实现

2.4.1 功能及局限

星表自动入库工具主要功能包括：

1. 为用户判断上载的 ReadMe 文件是否符合 ReadMe 文件标准，以及所选

入库的数据表在 ReadMe 文件中是否有相应的表结构（即能不能入库）。

2. 用户只要有相应权限的用户密码就可自动入库，服务器地址可以是本机也可是异地。

3. 用户可以同时录入星表中的多个数据表，即使这些数据表有不同的表结构，也能分别找到对应的表结构入库。

4. 能够自动处理不同赤经赤纬格式的情况，无论是度的形式，还是时分秒度分秒的形式。如果是度的形式就直接入库并加上索引；如果是时分秒度分秒的形式,就自动转换成度的形式并加进表里，再对度的形式的两列加索引。

星表 ReadMe 文件虽然有一定标准，但具体的星表其 ReadMe 文件还是有很多细节上的不统一，如赤经的格式。数据表的内容则更是复杂麻烦。有一些及其不规范的星表无法统一进来，用户需要修改 ReadMe 文件才能入库。

需要用户进行手动修改的情况：

1) 参数名出现重复或不符合规范的，参数名只能是字母、数字或下划线。
2) 参数名中出现“/”、“.”、“-”、“’”、“”，需要转换成下划线“_” 。
3) 赤经赤纬的单位不符合标准的，标准单位要求为：deg、h/min/s、deg/arcmin/arcsec。
4) 度的 Lable 标准是 RAdeg、DEdeg，时分秒度分秒的 Lable 标准是 Rah、Ram、RAs、DE-、DEd、DEm、DEs。如果有两个历元的时分秒度分秒格式的赤经赤纬，需要都按这个标准来写，而且还需要把 2000 历元写在 1950 历元的前面，工具会自动把前面的改成 RA_J2000、DE_J2000，把后面的改成 RA_B1950、DE_B1950。

2.4.2 程序流程图

下面我们通过示例介绍一下该工具的程序流程图（图 2-2），共分四个步骤：

(1) 首先用户上载 ReadMe 文件，该工具会自动判断是否符合 ReadMe 文件标准。如果是，则进入选择入库数据表的界面，界面中根据 ReadMe 文件提供的信息列出了该星表包含的文件供用户选择；如果不是，则需要重新上传。

(2) 接着用户选择需要入库的数据表，该工具自动判断 ReadMe 文件中是否有所选数据表的表结构（即所选数据表能不能入库），都能入库则该工具会解析所选数据表在 ReadMe 文件中对应的表结构信息，这一步很重要。另外用户可同时录入星表中的多个数据表，即使这些数据表有不同的表结构，该工具也能分别找到对应的表结构入库。

(3) 上载入库数据表文件。默认的数据表路径与星表 ReadMe 文件路径一样，如果路径不一样，用户可以重新选择路径。

(4) 最后进入入库选择界面。以第谷（Tycho）星表入库为例，如图 2-3所示，其中表名(tables name)的默认值是 ReadMe 文件提供的数据表表名，用户也可以修改表名，图中录入了 24 个数据表。用户设置好录入的数据库名(DataBase name)、数据库用户名(DB user)、数据库密码(DB Password)、数据库服务器地址(DB Server)、数据库端口号(DB Port)，其中数据库服务器地址可以是本机也可以是异地，用户只要有相应权限和数据库密码即可。图 2-3中的数据库服务器地址是本机。按 OK 键就开始自动入库，该工具会将解析的表结构信息自动生成 SQL 语句创建相应的表，并将数据表文件内容插入数据库。

入库结果如图 2-4所示，该示例是图 2-3中的表 trc_2，每一列对应星表的一项属性，每一行对应一个天体。该数据库的 RAdeg、DEdeg 是天体在天球上的位置坐标值（赤经、赤纬）。表的函数依赖关系很简单，赤经、赤纬决定其他参数，满足第三范式和 BC 范式，不需要再进行表的分解。另外，不同星表赤经、赤纬的格式可能不同，有的是十进制，有的是六十进制，该工具能自动识别不同的格式。

该工具用 Java 语言编写，开发环境配置如下：Eclipse3.0、jdk1.5.0_01、MySQL Server 4.1；运行环境配置如下：jdk1.5.0_01、MySQL Server 4.1。

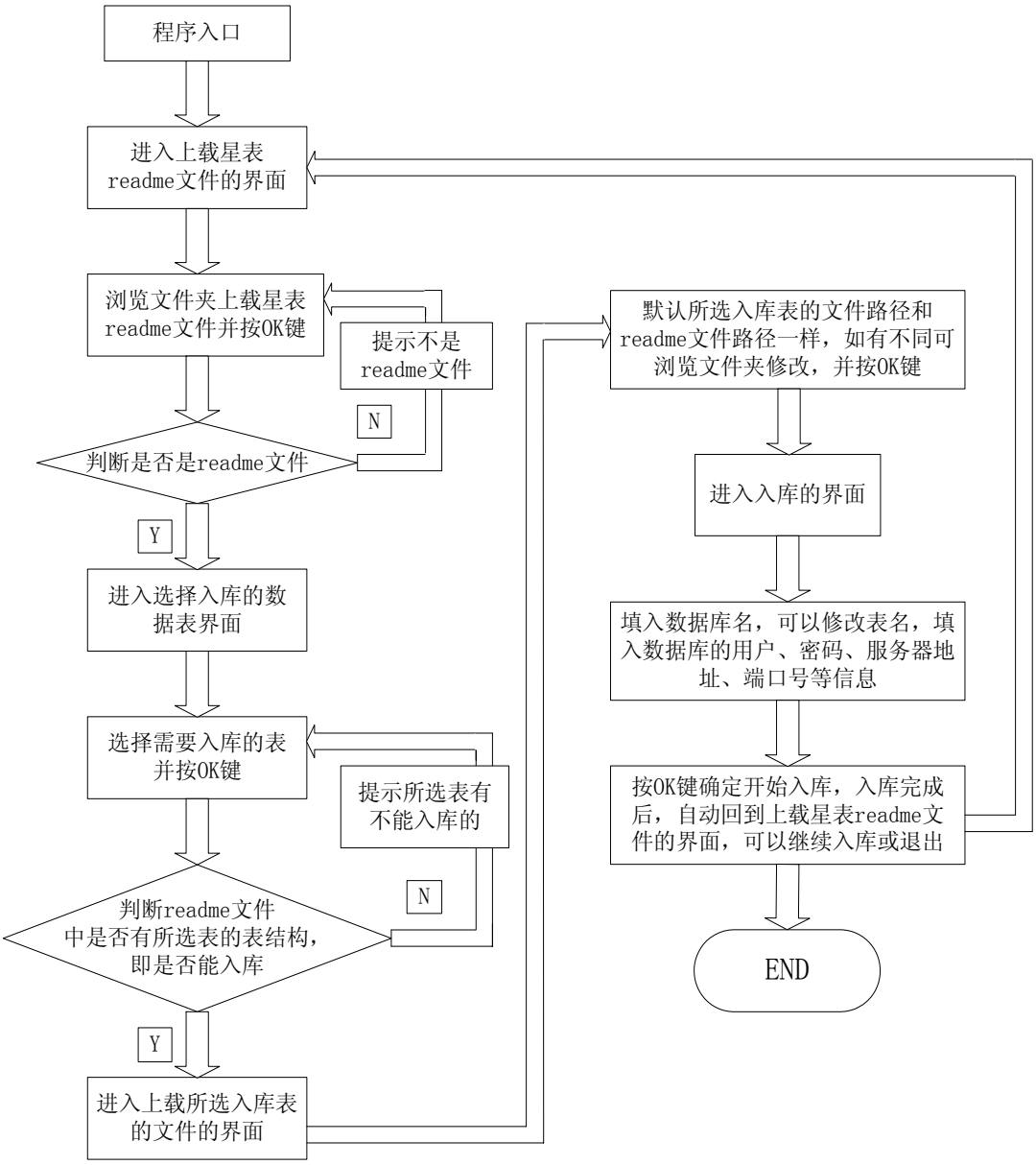


图 2-2 星表自动入库工具程序流程图

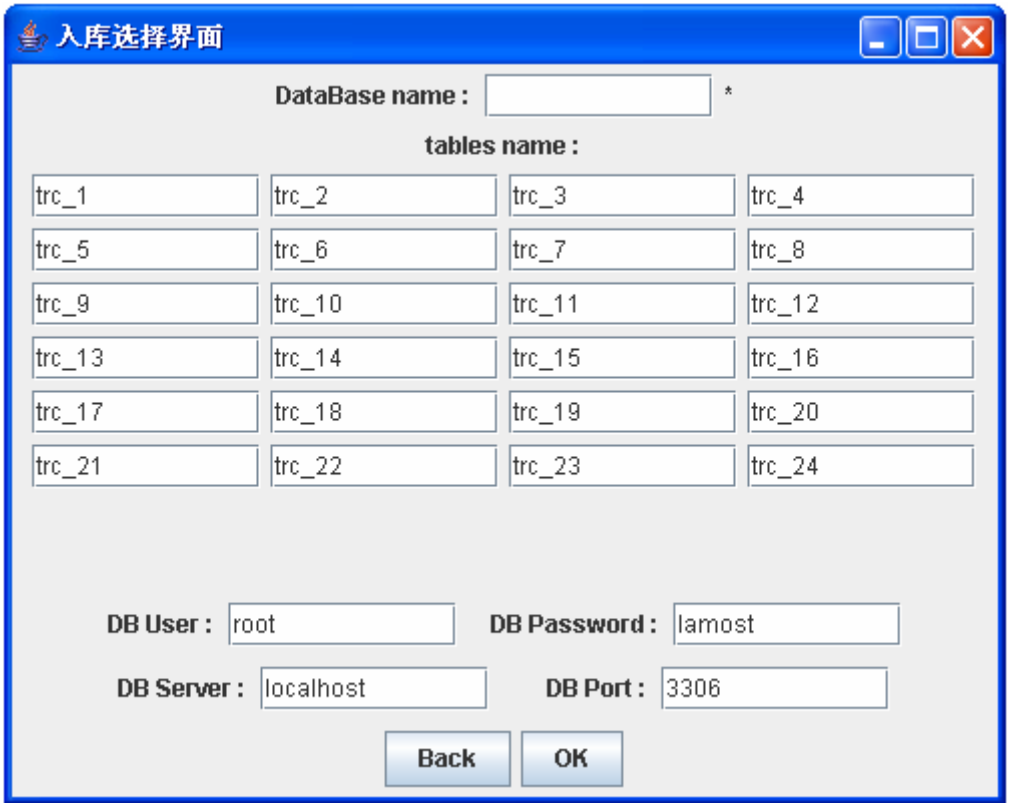
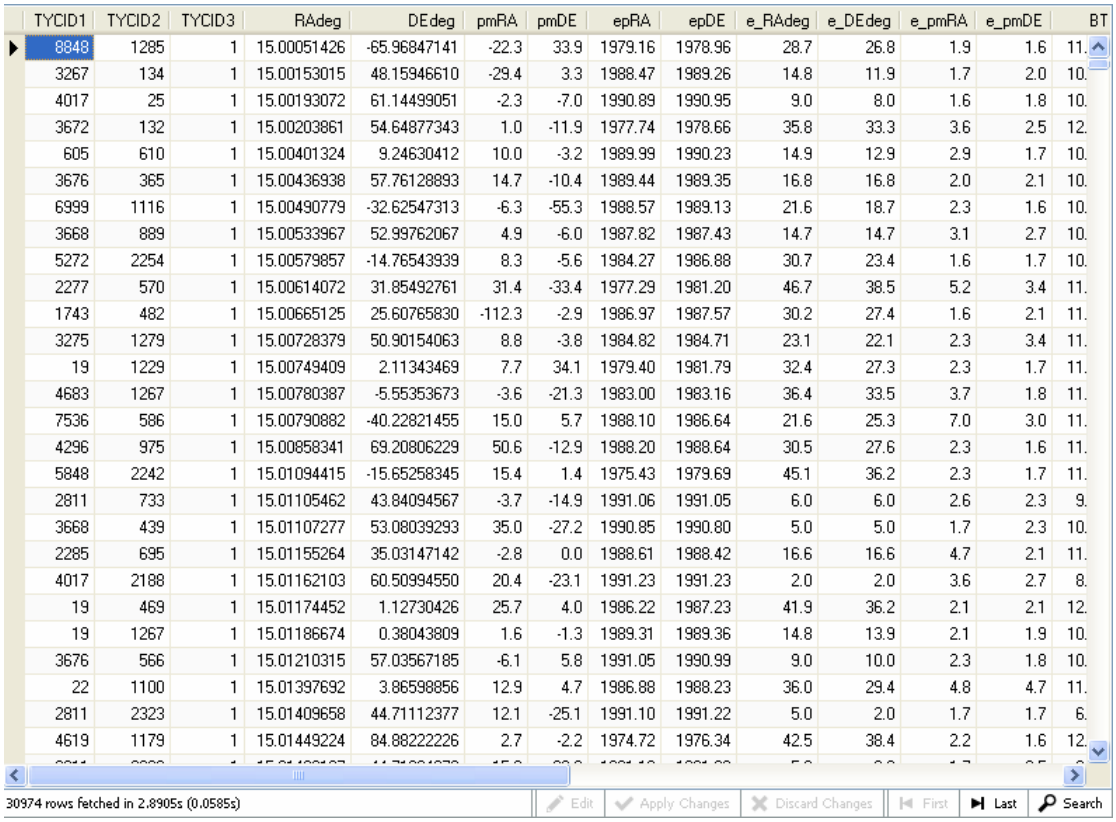


图 2-3 星表自动入库工具的入库选择界面



TYCID1	TYCID2	TYCID3	RAdeg	DEdeg	pmRA	pmDE	epRA	epDE	e_RAdeg	e_DEdeg	e_pmRA	e_pmDE	BT
8848	1285	1	15.00051426	-65.96847141	-22.3	33.9	1979.16	1978.96	28.7	26.8	1.9	1.6	11.
3267	134	1	15.00153015	48.15946610	-29.4	3.3	1988.47	1989.26	14.8	11.9	1.7	2.0	10.
4017	25	1	15.00193072	61.14499051	-2.3	-7.0	1990.89	1990.95	9.0	8.0	1.6	1.8	10.
3672	132	1	15.00203861	54.64877343	1.0	-11.9	1977.74	1978.66	35.8	33.3	3.6	2.5	12.
605	610	1	15.00401324	9.24630412	10.0	-3.2	1989.99	1990.23	14.9	12.9	2.9	1.7	10.
3676	365	1	15.00436938	57.76128893	14.7	-10.4	1989.44	1989.35	16.8	16.8	2.0	2.1	10.
6999	1116	1	15.00490779	-32.62547313	-6.3	-55.3	1988.57	1989.13	21.6	18.7	2.3	1.6	10.
3668	889	1	15.00533967	52.99762067	4.9	-6.0	1987.82	1987.43	14.7	14.7	3.1	2.7	10.
5272	2254	1	15.00579857	-14.76543939	8.3	-5.6	1984.27	1986.88	30.7	23.4	1.6	1.7	10.
2277	570	1	15.00614072	31.85492761	31.4	-33.4	1977.29	1981.20	46.7	38.5	5.2	3.4	11.
1743	482	1	15.00665125	25.60765830	-112.3	-2.9	1986.97	1987.57	30.2	27.4	1.6	2.1	11.
3275	1279	1	15.00728379	50.90154063	8.8	-3.8	1984.82	1984.71	23.1	22.1	2.3	3.4	11.
19	1229	1	15.00749409	2.11343469	7.7	34.1	1979.40	1981.79	32.4	27.3	2.3	1.7	11.
4683	1267	1	15.00780387	-5.55353673	-3.6	-21.3	1983.00	1983.16	36.4	33.5	3.7	1.8	11.
7536	586	1	15.00790882	-40.22821455	15.0	5.7	1988.10	1986.64	21.6	25.3	7.0	3.0	11.
4296	975	1	15.00858341	69.20806229	50.6	-12.9	1988.20	1988.64	30.5	27.6	2.3	1.6	11.
5848	2242	1	15.01094415	-15.65258345	15.4	1.4	1975.43	1979.69	45.1	36.2	2.3	1.7	11.
2811	733	1	15.01105462	43.84094567	-3.7	-14.9	1991.06	1991.05	6.0	6.0	2.6	2.3	9.
3668	439	1	15.01107277	53.08039293	35.0	-27.2	1990.85	1990.80	5.0	5.0	1.7	2.3	10.
2285	695	1	15.01155264	35.03147142	-2.8	0.0	1988.61	1988.42	16.6	16.6	4.7	2.1	11.
4017	2188	1	15.01162103	60.50994550	20.4	-23.1	1991.23	1991.23	2.0	2.0	3.6	2.7	8.
19	469	1	15.01174452	1.12730426	25.7	4.0	1986.22	1987.23	41.9	36.2	2.1	2.1	12.
19	1267	1	15.01186674	0.38043809	1.6	-1.3	1989.31	1989.36	14.8	13.9	2.1	1.9	10.
3676	566	1	15.01210315	57.03567185	-6.1	5.8	1991.05	1990.99	9.0	10.0	2.3	1.8	10.
22	1100	1	15.01397692	3.86598856	12.9	4.7	1986.88	1988.23	36.0	29.4	4.8	4.7	11.
2811	2323	1	15.01409658	44.71112377	12.1	-25.1	1991.10	1991.22	5.0	2.0	1.7	1.7	6.
4619	1179	1	15.01449224	84.88222226	2.7	-2.2	1974.72	1976.34	42.5	38.4	2.2	1.6	12.

图 2-4 录入的数据库示例

2.5 星表搜索服务

基于星表数据的自动入库工具，我们将 CDS VizieR 所有星表的 ReadMe 文件中描述的元数据信息提取出来，录入到天文星表元数据库中，供世界数据中心天文中心提供的星表搜索服务查询使用。该星表搜索服务网址：http://badc.lamost.org/cnweb/modules/catsquery/。服务界面如图 2-5所示，举例搜索与 SDSS 相关的星表，如图 2-6所示，共有 42 个星表。

元数据库的建表 SQL 语句如下：

```
create table badc.ReadMeV3 (badcID int(11) not null auto_increment primary key ,Designation varchar(80) not null ,AbbreviatedTitle varchar(80) not null ,FullTitle text not null ,Authors text not null ,Reference text not null ,BibCodes varchar(64) not null ,keywords text
```

```
null ,Description text null ,Objects text null ,FileSummary
text null ,SeeAlso text null ,NomenclatureNotes text
null ,ByteByByteDescription text null ,Note text null ,Others
text null ,Updated date not null);
```

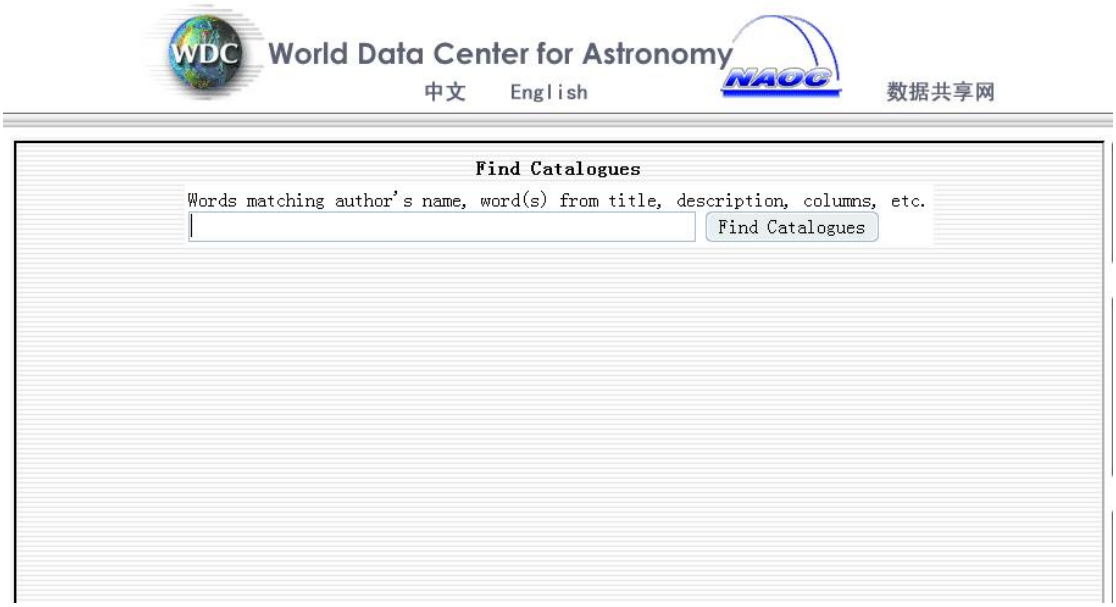


图 2-5 星表搜索服务界面

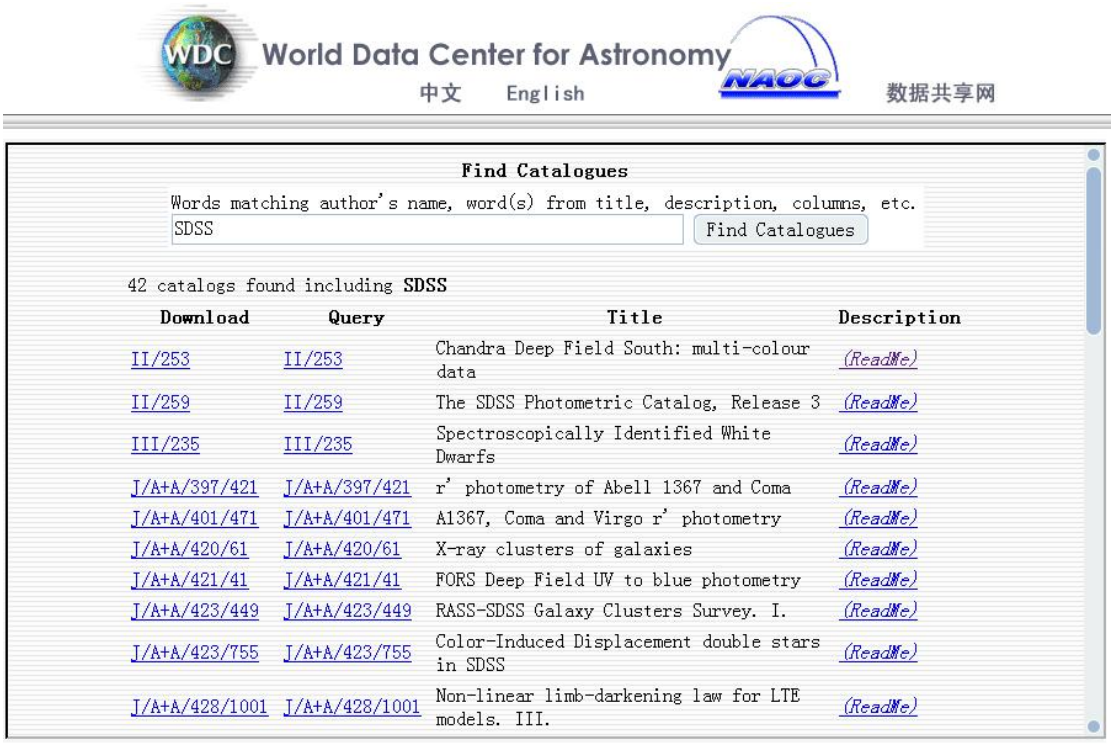


图 2-6 星表搜索服务示例

2.6 小结

本章首先介绍了天文星表数据库，然后着重探讨了星表 ReadMe 文件标准、数据文件格式问题、如何对星表数据统一进行预处理、以及基于星表 ReadMe 文件的星表数据自动入库工具的实现，最后介绍了如何建立天文星表元数据库和星表搜索服务。这部分工作是进一步进行数据融合和数据挖掘工作的基础。

第 3 章　天文数据融合算法的研究和工具的开发

天文数据融合的核心是基于不同星表的天体在位置上的相关性，来进行星表间的交叉证认工作。将不同波段的星表、特别是巡天项目产生的大型星表，进行交叉证认，一直是很多天文学研究的基础，尤其对从事多波段数据挖掘和统计分析工作的研究人员具有重要意义。本章的天文数据融合研究是在上一章星表自动入库和统一预处理的基础上进行的。为了提高数据查询和处理的效率，需要研究高效的数据库索引算法。对比了多种索引算法的优劣，并首次提出了一种优化的交叉证认算法。而且一个个小程序的应用势必增加天文学家学习与使用的难度和繁琐性，为此我们开发了高度自动化的工具来克服这些困难。

3.1　交叉证认概述

3.1.1　交叉证认定义

交叉证认是指两个不同来源的星表数据，以一个星表源（天体）的位置为中心，对于该星表中的每一个源，都在另一个星表中寻找对应体。若两个源之间的角距离满足一定的条件，就认为这两个源极有可能是对应体。如果是来自两个不同波段的星表 A 和星表 B，通过交叉证认，可以找到星表 A 的源在另一个波段星表 B 的对应体，从而获得该源在两个不同波段的属性。相应地可以得到一个新的多波段星表，它的每一行包括星表 A 的所有参数、星表 B 的所有参数以及两个对应源之间的角距离。经过来自多个波段的星表的交叉证认，我们就获得了多波段数据。因此，通过交叉证认能够获得同一目标源在多个波段的属性，故多波段数据融合是加深对天体的认识和促进发现新的天体或现象的革命性的一步，也为进一步的统计分析和数据挖掘打下基础。

由于不同望远镜观测精度的差异、位置误差的不同，使得交叉证认的对应体

结果可能有以下几种不同情况：一对一、一对多、多对一、一对无、无对一等。（如图 3-1所示）对于一对一的源，我们可以基本上认为该对应体就是同一个源；对于一对多、多对一的源，就需要进一步运用概率的方法来确定。目前很多是取角距离最近的源为对应体或最亮的源为对应体。为有效地选取对应体，还可以利用星表中的其它信息，例如在 USNO-A2.0 和 USNO-B1.0 种给出 B 和 R 的星等，而 B-R 就是一个很好的参数用以区分最终的对应体。得到对应体后，我们就可以进一步做统计研究和数据挖掘工作。对于一对无或无对一的情况，就需要天文学家做专门的天文学上的解释。

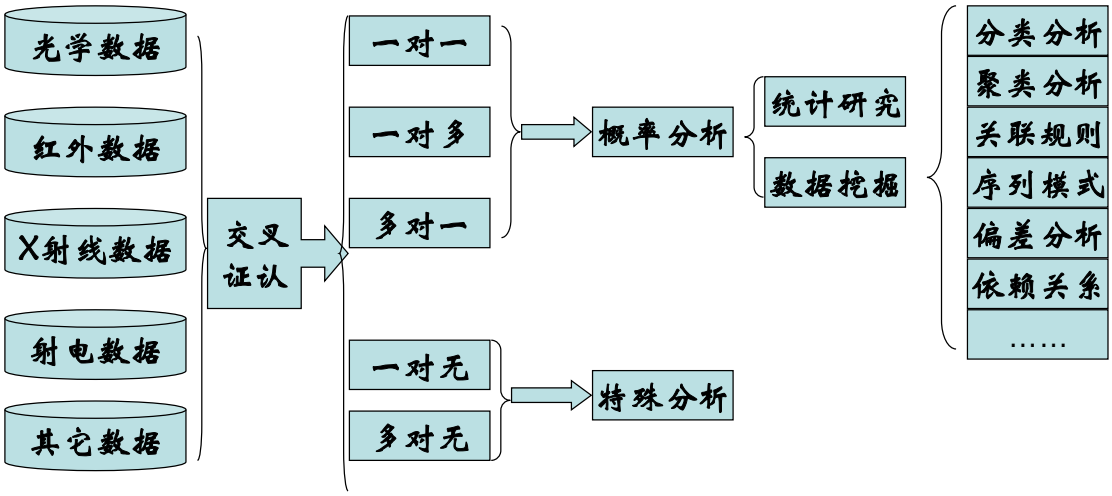


图 3-1 多波段数据分析流程图

3.1.2 国内外的研究情况

目前，国外现有的交叉证认工具有：VizieR、Simbad、Aladin、MAST、NED、OpenSkyQuery、SDSS catsjob、Topcat 等。

VizieR[3] 是 CDS 开发的数据融合工具。它已经收集了 6000 多个星表，其数据在不断更新。它提供了目前已发表数据的各种查询方式。对单个星表查询，用户可以自由选择所需参数。另外它也提供了简单的交叉证认功能，用户可以上

传文件与感兴趣的星表交叉证认，但结果还需要用户二次加工。

Simbad[23] 提供查询服务，查询功能及小样本多源查询（即小样本交叉证认）功能类似 VizieR，主要提供点源的证认情况。选源时可以通过 Simbad 查看所选源的类型及是否已被证认过等信息。

Aladin[4] 是 CDS 开发的数据整合工具，提供查询、交叉证认、光谱分析等功能。它在天体的多波段交叉证认、观测准备和新数据集的质量控制等方面显示出优越性，但任意海量数据交叉证认还没有实现。

MAST（The Multimission Archive at STScI，简称 MAST）[24] 是 NASA 资助的基金项目，为天文学界提供各种各样的天文数据库的支持和服务，主要用于获得可见光、紫外和近红外的相关的光谱数据。也提供了仅限于小样本的交叉证认。

NED[25] 提供了批处理查询功能，可以查询某源在其收录资料中的所有信息，包括最新发表文章中的数据。功能上类似 Simbad，NED 主要是河外源的认证。但对多源的查询，它不是在线查询而是通过邮件来进行查询，并且只限于小样本的证认服务。

OpenSkyQuery[5] 是美国虚拟天文台开发的交叉证认工具。它用一种通用的天文数据查询语言 ADQL 实现了数据查询、交叉证认等功能，用户可以选择交叉的星表，也可以上传文件与选定的星表交叉证认。其服务页面提供了各种数据库、交叉证认算法的模版、用户指导等。但该系统查询或证认的行数不能超过 5000 行。

SDSS CasJobs[26] 是一个用于大型科学星表的在线工作平台，其设计为在 web 环境中模拟并加强本地的自由格式查询访问。它仅提供了 SDSS 巡天数据的 SQL 语言查询和数据下载服务，并且每次查询的结果限制在 500M 之内。

Topcat[27] 是一个可以互动地将数据表图形化的浏览器和编辑器。它提供了各种方法浏览和分析数据表，包括浏览核心数据、表的原始信息和列的元数据、以及画图工具、统计计算、交叉证认等，但内存一般限制在 256M 范围内。

由上面的服务可以看出每个数据中心的服务都有其局限性，如数据资源局限

和功能局限，因此目前国外这些大数据中心提供的数据获取或交叉证认服务还远远不能满足进行大样本工作的天文学家的需求。

综上所述，对目前存在服务的优点总结如下：

(1) 用户界面非常友好、可查询、可视化等功能齐全；

(2) 支持的输出格式齐全（如文本、网页、VOTable 等），更新速度快；

同时，这些服务具有如下缺点：

(1) 不能进行海量数据融合；

(2) 数据资源局限，只能与特定的数据交叉证认；

(3) 用户不能从一个服务中找到任意需要的星表；

(4) 参数不能自由选择；

(5) 交叉证认结果需要用户进一步加工；

(6) 没有对证认的结果进行分类。

3.1.3 交叉证认的步骤和技术难点分析

交叉证认的步骤如下：

(1) 输入：两个星表 A、B（星表 A：包含那些需要寻找对应体的源；星表 B：在该星表中寻找对应体）；

(2) 若两个源之间的角距离 d 满足: d≤d_max，则这两个源被认为是匹配的对应体；

(3) 输出：一个新的多波段星表，它的每一行包括星表 A 的所有参数、星表 B 的所有参数、两个对应源之间的角距离；

(4) 交叉证认半径满足下式[28] ：

$$d \le 3\sqrt{r_1^2 + r_2^2}$$

其中 d 是两个源之间的角距离，r1 和 r2 为两个星表的误差半径。

严格而言，角距离 d 应该是两点之间的球面角距离。考虑到大数据量运行的效率，角距离 d 在很小时，可以取如下的近似计算公式：

$$d = \sqrt{((\mathrm{RA\_A} - \mathrm{RA\_B})\cos((\mathrm{DEC\_A} + \mathrm{DEC\_B})/2))^2 + (\mathrm{DEC\_A} - \mathrm{DEC\_B})^2}$$

在实现海量多波段数据交叉证认的过程中存在一些技术难点：

(1) 数据量大，数据复杂且高度分布

数据资源高度分布，收集数据资源的能力有限；天文数据来自于不同的观测仪器、观测地点和观测时间；数据形态和格式复杂，即使是按标准的 ReadMe 文件也都各有不同。需要考虑诸多繁琐因素，因此大数据量的异地传输不现实。

(2) 仪器灵敏度、分辨率和误差精度的不同

这些因素进一步导致了数据的复杂性，这为交叉证认的实施和结果的分析带来相当高的难度，且用户的精度要求很难满足。

(3) 天文数据处理程序的时间复杂度高、占用内存大

理论上交叉证认的时间复杂度是 N×N，再加上异地传输时间，所消耗的时间非常长，而且内存常常成为瓶颈。因此，实现海量天文数据的交叉证认需要对算法进行研究，以降低时间复杂度和突破内存的限制。

如果用户想交叉证认自己的一个大数据量星表和服务器提供的一个星表，需要解决大数据量异地传输的问题。为此，我们提出的解决方案是：因为交叉证认的过程只需要用到赤经赤纬，可先交叉证认出对应星表的名字标识，再用 SQL 语句调用对应的其他参数，显示结果。具体步骤如下：

(1) 提取只有赤经、赤纬、ID 三个属性的表；

(2) 把该表以文本格式上传；

(3) 将上传文件在服务器端入库；

(4) 交叉证认；

(5) 根据对应结果再调对应的其他参数；

(6) 显示结果。

3.2 交叉证认算法的研究

当对数据库进行查询等操作时，常规的办法需要遍历整个数据表。这在数据量巨大的时候是一项非常耗时的工作。例如，对 TB 量级的数据，需要一天或者

更多的时间来进行查询。而索引可以大大提高 DBMS（Database Management System，间称 DBMS）的检索性能，一个高效的索引能够使数据库在 2-3 个寻道时间（2-3 毫秒）内找到任何记录。

海量天文星表的交叉证认是通过赤经和赤纬这二维空间天体坐标位置来进行的，其涉及的数据库非常大而且结构复杂，如果仅对赤经和赤纬进行索引，不能从根本上改善检索的性能。要从根本上提高空间检索的性能必须实现空间坐标二维索引。目前在 DBMS 中对海量天文数据进行索引主要有三种实现方式：第一种是对赤经赤纬分别建一维 B-tree 索引，这是非常成熟的数据库索引方法；第二种是建立空间二维索引：如 R-tree、G-tree、kdB-tree、PK-tree、SR-tree 等，Clive Page 对许多数据库系统的空间索引进行了研究[29] ，发现这些索引功能的性能都不尽人意；第三种是利用某种映射关系将二维的空间参量映射到一维参数空间，把赤经赤纬合二为一，降低检索维度，这可以称为伪空间索引。目前较好的两个方法是 HTM[30] 和 HEALPix[31] 。

为了解决大数据量交叉证认的问题，我们研究了 B-tree 和 HTM 两种索引算法来进行交叉证认。第一种算法是在数据库中建立 B-tree 索引，提高数据库操作的性能。这种算法证认精度高，但只能用来处理小样本的交叉证认问题，当星表的行数大于几十万行时会遇到内存瓶颈。第二种算法是利用 HTM 空间索引替代两个 B-tree 索引，交叉证认过程变成了数据库的自然连接，解决了内存限制的问题，但此算法证认精度低，漏源概率大。所以，这两种算法还不能很好地满足海量数据交叉证认的需求。

随后，我们借鉴了以上两个算法的优点，并引入了 kd-tree 算法找最近邻的思想，创新地提出了 HTM 索引分区与 kd-tree 找最近邻算法，彻底解决了内存限制和算法复杂度高的问题，并保证了高的证认精度。下面我们将分别介绍这几种方法的工作原理。

3.2.1 基于 B-tree 索引的交叉证认

平衡二叉树（B-tree）一维索引与 DBMS 中的查询优化器实现了完美的结合。

几乎所有的 DBMS 都使用 B-tree 作为默认的索引方式。大量的实践证明一维 B-tree 索引是非常成功的。与其他的空间索引方法相比（如 R-tree、G-tree、kdB-tree、PK-tree、SR-tree 等），一维 B-tree 是较好的索引方法。它提供快捷的数据插入、删除、查找功能，而且能够保持平衡结构，并具有高空间利用率的优点。图 3-2给出了一个 B-tree 索引的应用实例。

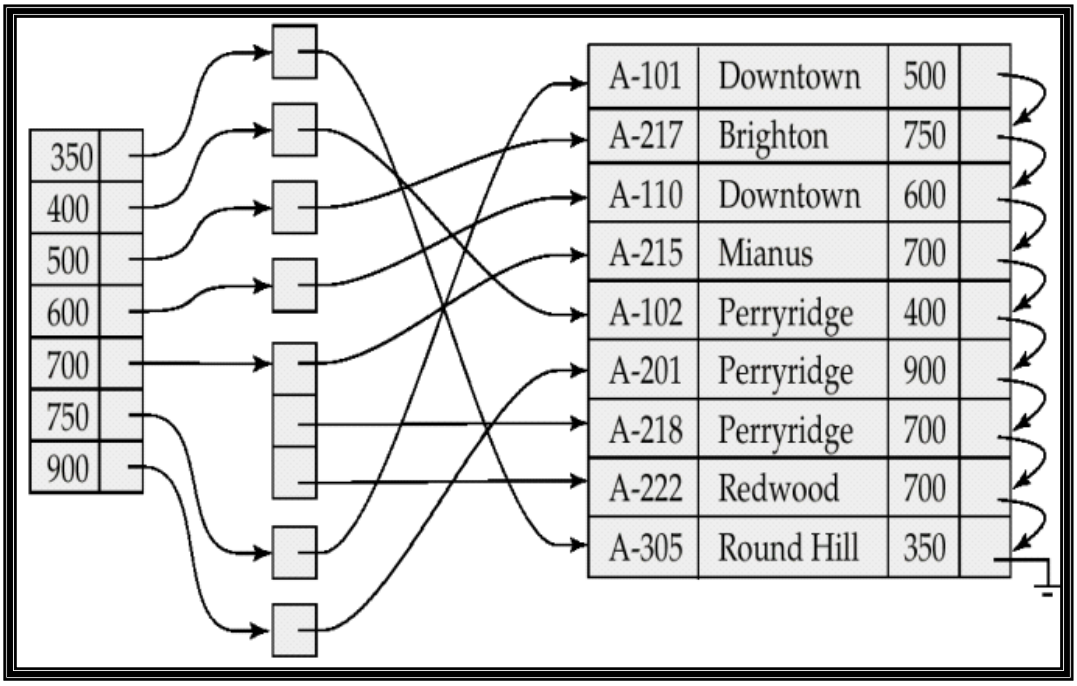


图 3-2 B-tree 索引

为了提高查询和交叉证认的速度，我们分别为两个星表数据库的赤径、赤纬两列建 B-tree 索引。图 3-3和图 3-4显示了用 B-tree 索引交叉证认的过程。此算法以小表为中心，在大表中遍历寻找证认源，并用证认判断公式判断是否证认源。该方法的证认精度高，算法的复杂度为 $O(N^2)$，速度慢，并且内存存在限制。

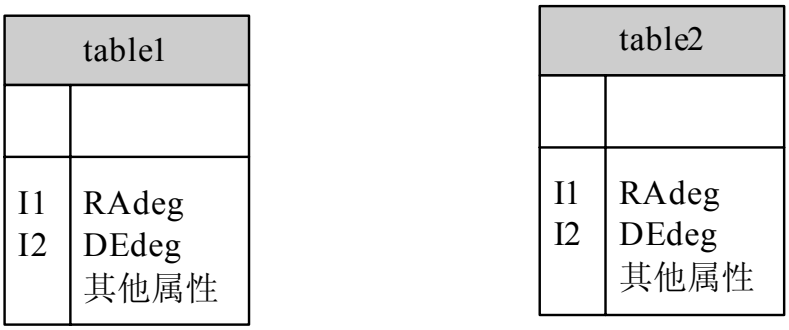


图 3-3 基于 B-tree 索引的交叉证认（1）

图 3-4 基于 B-tree 索引的交叉证认（2）

3.2.2 基于 HTM 索引的交叉证认

为了解决内存限制的问题，可以采用降维索引方法，目前有两种比较好的候选方案，即多级三角划分法（Hierarchical Triangular Mesh，简称 HTM）[30] 和多级等面积同纬度划分法（Hierarchical Equal Area isoLatitude Pixelisation，简称 HEALPix）[31] 。它们的基本思路大致相同，都采用天区分割方式，对整个天区进行迭代式的层层划分。然后按照一定的编码规则给每个子天区单元编号，我们称其为 pcode。这个 pcode 便与赤经、赤纬产生了一定的对应关系，从而实现了二维空间到一维空间的映射。数据库系统便可以采用成熟的索引技术对 pcode 进

行索引和操作，从而实现快速检索的目的。

我们采用了 HTM 索引的交叉证认算法。HTM 是由约翰·霍布金斯大学的 Alex Szalay、Peter Kunszt、Ani Thakar 设计的一种天区划分方案，首先应用于 SDSS 巡天数据。起始状态将整个天区分为 8 等份，上下各四个球面直角三角形。然后以每个球面三角形各边中点为新的顶点对原三角形四等分，依此类推。0 到 5 级的划分情况如图 3-5所示。HTM 的编码方式如图 3-6和图 3-7所示。最顶层 8 块按顺时针方向依次命名为 N0、N1、N2、N3、S0、S1、S2、S3。由于都是四等份，父单元与子单元编码长度相差两个比特。HTM 已经被空间望远镜科学研究所（STScI）用来索引 GSC-II 星表。欧洲空间局（ESA）的 GAIA 项目也计划采用 HTM 对数据进行索引。

如何选择合适的划分级次，以便每个单元中源的个数不会太多，同时单元面积要大于源的位置误差范围。表 3-1给出了 HTM 方案中不同划分级次对应的单元数和单元大小。为了使用 pcode，不得不对现有的数据表进行结构上的调整。要为每个表中增加一列以存储相应的 pcode，这需要一定的工作量。

HTM 算法根据星表的精度选取一定的 HTM 级数，再由星表的赤径、赤纬计算出对应 HTM 索引的 pcode 值。基于 HTM 索引的交叉证认过程如图 3-8所示。该算法与 B-tree 索引算法相比，是将两个 B-tree 索引变成一个 HTM 空间索引。这样交叉证认过程也由原来的 N×N 遍历变成了数据库的自然连接。但此算法仍然没有解决算法复杂度的问题。由于没有证认的判断，只是要求两个星表级数一样，即要求两个表精度差不多，所以它的认证精度低，而且还具有漏源概率高，存在大量一对多、多对一对应体混杂的缺点。

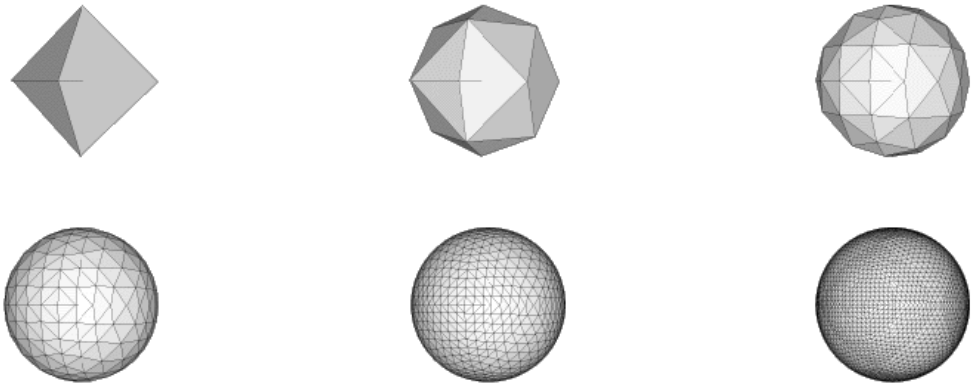


图 3-5　HTM 算法天区划分

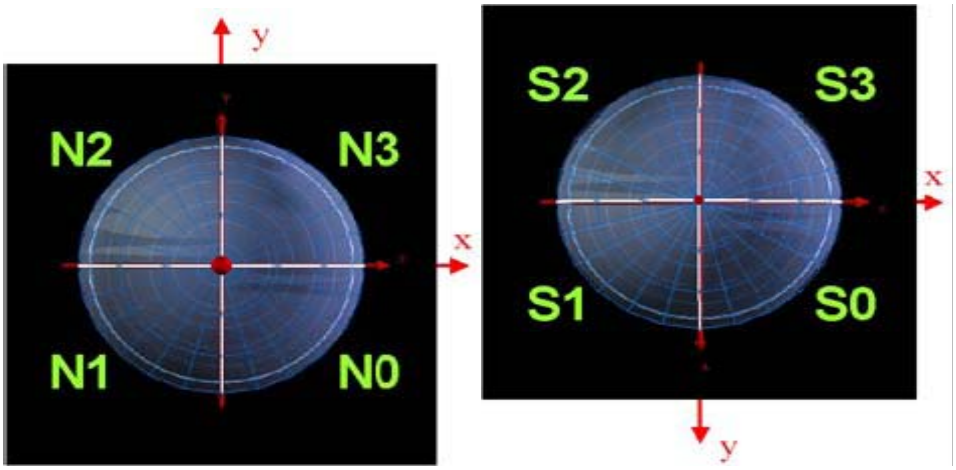


图 3-6　HTM 算法编码方案（1）

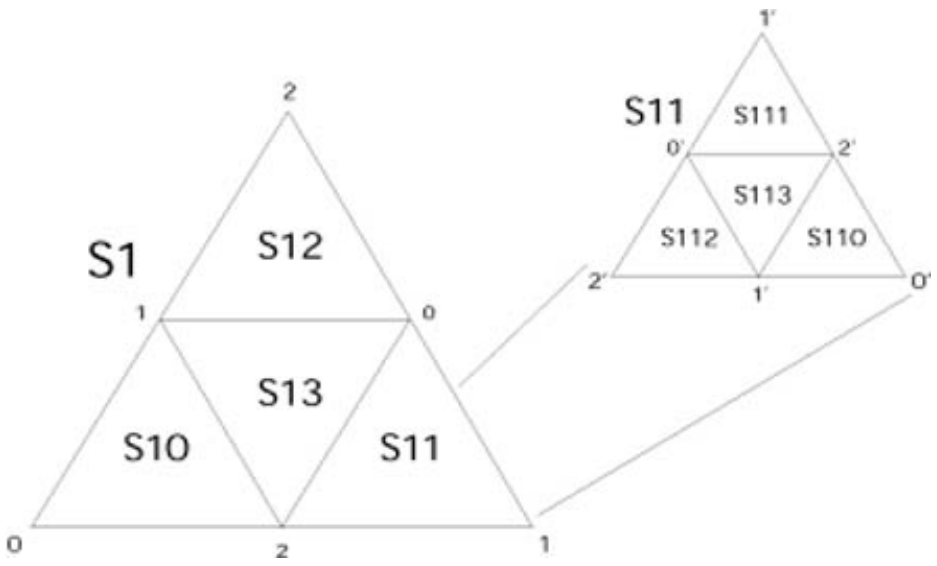


图 3-7 HTM 算法编码方案（2）

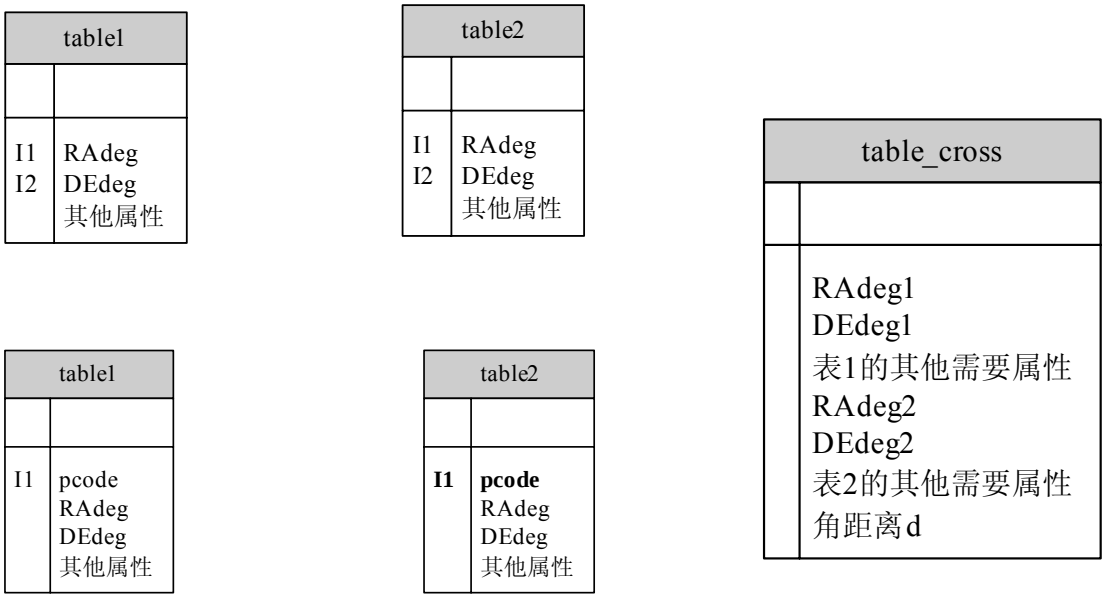


图 3-8 基于 HTM 索引的交叉证认

表 3-1 HTM 不同划分级次对应的单元数和单元大小

级次	面积 (arcsec^2)	单元数
10	1.77E1	8,388,608
11	4.43E0	33,554,432
12	1.11E0	134,217,728
13	2.77E-1	536,870,912
14	6.92E-2	2,147,483,648
15	1.73E-2	8,589,934,592
20	1.69E-5	8,796,093,022,208
25	1.65E-8	9,007,199,254,740,922

3.2.3 基于 HTM 索引分区与 kd-tree 找最近邻算法的交叉证认

为了彻底解决内存限制和算法复杂度高的问题，我们借鉴了以上两个算法的优点，并引入了 kd-tree 算法找最近邻的思想，创新地提出了这个新的方案。算法以小表为中心，把 HTM 索引当作分区，接着分别对每个分区的大表，把赤径、赤纬作为参数建 kd-tree，并对该分区内小表的每个源，在 kd-tree 中找最近邻，算法用证认判断公式判断是否证认。我们需要为每个表建立一个主键 id_htm，并建立一个新表只记录 id_htm 和 pcode 列信息。由于 HTM 级数选得比较小，比星表精度低，漏源的可能性小。此算法的优点是用空间索引解决了内存限制问题，通过分区和找最近邻降低了算法复杂度，提高了速度和证认精度。

3.3 交叉证认工具的开发和应用

通过上节对提高交叉证认效率的算法研究，设计出了优化的策略，这为进一步的交叉证认工具的开发打下了坚实的基础。我们开发的交叉证认工具主要实现了两个服务：一个是服务器端两星表交叉证认；另一个是用户上传星表与服务器端星表交叉证认。在前者中两星表直接进行交叉证认，后者则先将本地数据自动入库再进行交叉证认。这个交叉证认工具的核心是自动入库模块和交叉证认模

块，它们实现了主要的自动入库和交叉证认功能，这两个模块的具体细节将在本节中详细说明。程序还实现了对交叉证认结果的分类和参数的自由选择等功能，以及与可视化工具 VOPlot 的集成。该工具以客户端/服务器模式的 web 网页形式发布。此工具为多波段数据融合提供了便利，是对两个大星表交叉证认工作的预研究。

3.3.1 功能和服务

本交叉证认工具的功能和服务描述如下：

(1) 实现大数据量星表交叉证认

解决了一直以来存在的大数据量星表本地、异地交叉证认困难的问题。大数据量有助于排除由于样本小带来的泊松误差的影响，促进有趣天体或现象的发现的概率。比如某些概率是百万分之一或一亿分之一量级的未知或稀有天体或天文现象，则有可能在大样本中发现。

(2) 用户可以任意选择两个星表交叉证认

解决了数据资源局限性的问题，实现了本地和异地的数据的交叉证认。如果用户想交叉证认服务器端的两个星表，那么证认可以在服务器端进行，而把最终结果传给用户。若用户要使用自己的星表与服务器的星表进行交叉证认，只需上传自己星表的元数据（ReadMe）文件和数据文件，在服务器端利用自动入库模块就可以实现星表的自动入库功能，从而可以方便地交叉证认。

(3) 用户可自由选择输出参数

为了节省存储空间，减少传输量，无须将两个星表的所有参数都传给用户，而只要把用户需要的参数输出即可。

(4) 用户可对交叉证认的结果选择

可以任意选取一对一、一对多、一对无、无对一的情况或这几种结果的任意组合。

(5) 证认结果无需二次加工

直接给出用户需要的结果，每一行包括用户选择的两个星表参数和对应源的

角距离的多波段星表。

(6) 用户可以选择交叉证认的两个星表的误差半径

对于所有源误差半径都一样的星表，可选择误差半径为一常数；而对于星表中每个源有其相应的误差半径，并用星表的一列数据来表示的情况，可以选择星表相应的列名为误差半径。

(7) 输出格式的多样化

支持 VOTable、ASCII、CSV、HTML 等格式。

(8) 与 VOPlot 可视化工具的集成

VOPlot 是印度虚拟天文台开发的 Java 程序，可以画不同的图（散点图、直方图、带误差棒的图），数据格式为 VOTable。可以对任何星表作图，而且图可以以 eps 格式保存。与交叉证认工具集成后，用户可以对交叉证认后的数据进行加工，例如增加列或进行一些简单的运算生成新列。用户也可以通过 VOPlot 来可视化数据，如选择已有的或新计算的参数来做散点图、直方图等。

3.3.2 服务界面

下面以 X 射线波段的 ROSAT 星表和光学波段的 TYCHO 星表做交叉证认为例，说明我们开发的交叉证认工具的功能。ROSAT 星表的误差半径选择了 PosErr 列，TYCHO 星表的误差半径选择了 5 角秒。本工具主要实现了两个服务：一个是服务器端两星表交叉证认服务，如图 3-9界面中上部的表格所示；另一个是用户上传星表与服务器端星表交叉证认服务，如图 3-9界面中下部的表格所示。用户要使用某服务只需填写相应的表格提交即可。图 3-10展示的是此工具的交叉证认结果的分类和参数的自由选择等功能，用户选择需要的参数和交叉证认的分类结果即可。

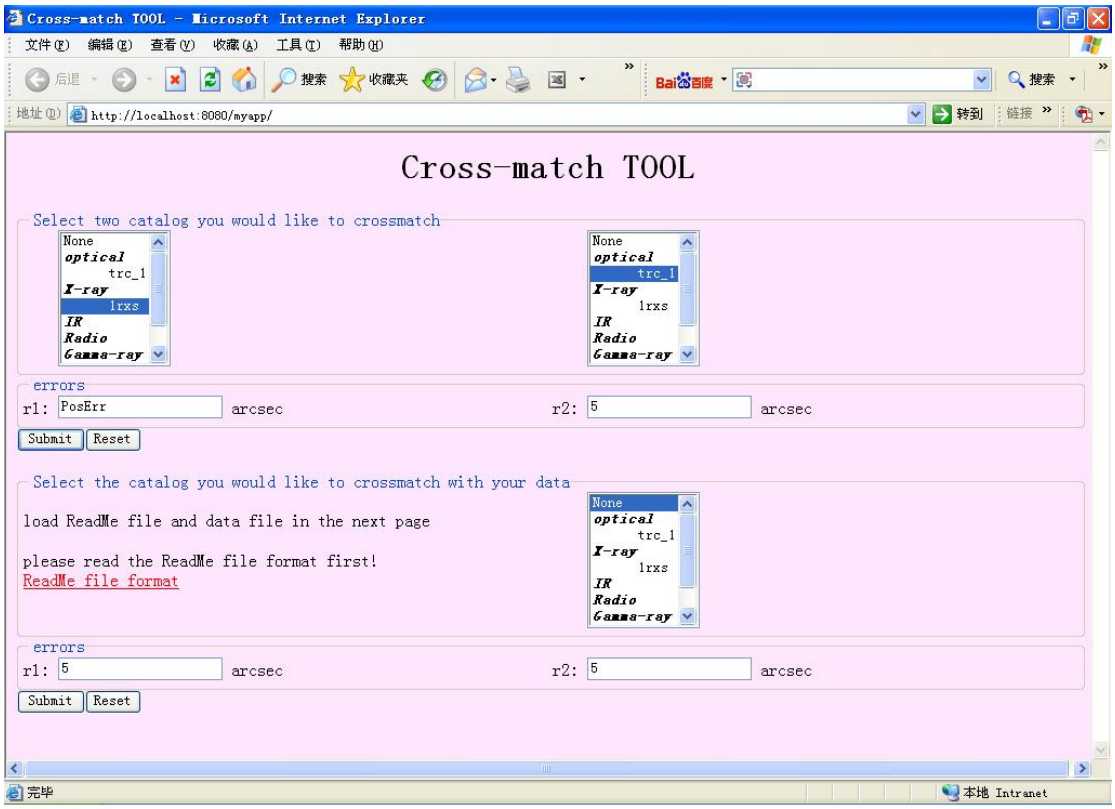


图 3-9 交叉证认工具主页面

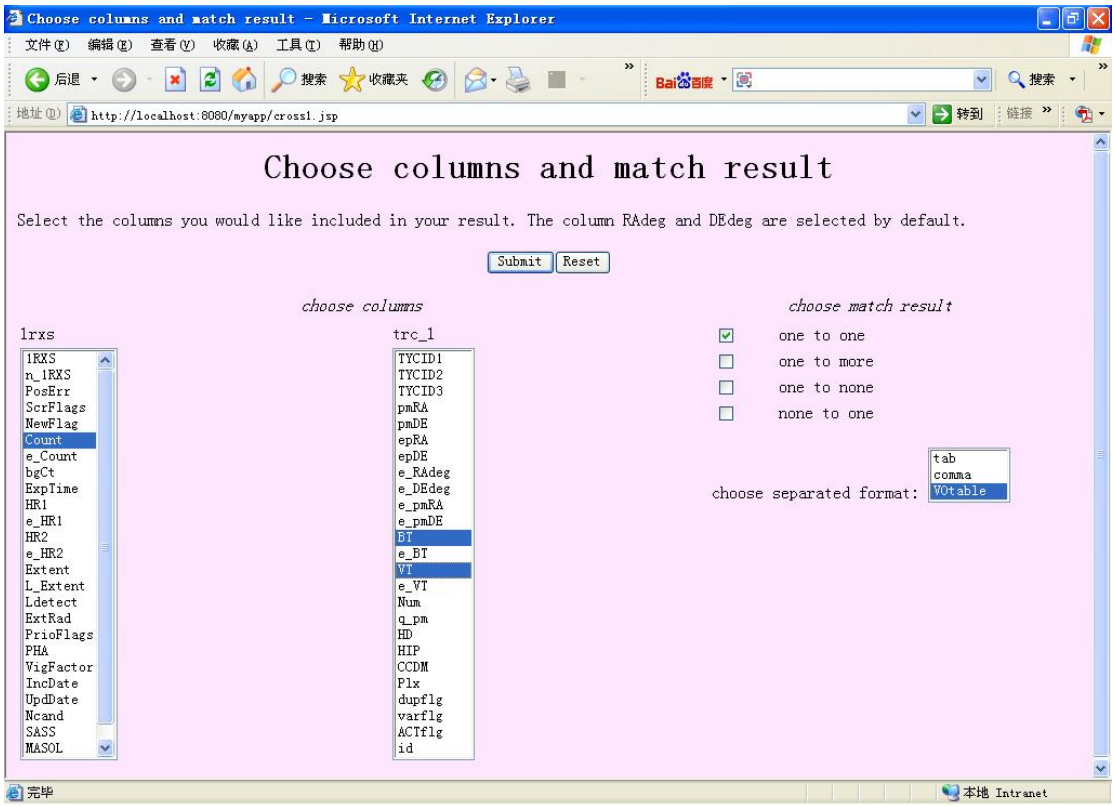


图 3-10　选择参数和交叉证认结果页面

3.3.3　工具程序流程图

如图 3-11，进入交叉证认主页面后（图 3-9），有两个服务对应的两条流程。一是交叉证认服务器上两个星表：用户选择星表和误差半径，提交表格；工具就会判断填写信息是否完整，若完整，则进入选择参数和交叉证认结果的页面（图 3-10）；用户选择后，工具就调用交叉证认模块进行证认；最后，VOPlot 输出可视化结果（图 3-13）及下载结果（图 3-12）。第二个服务是用户上传自己的星表与服务器上的星表交叉证认，其与第一个服务的区别是中间多了一个上载文件页面（图 3-14）和自动入库模块。

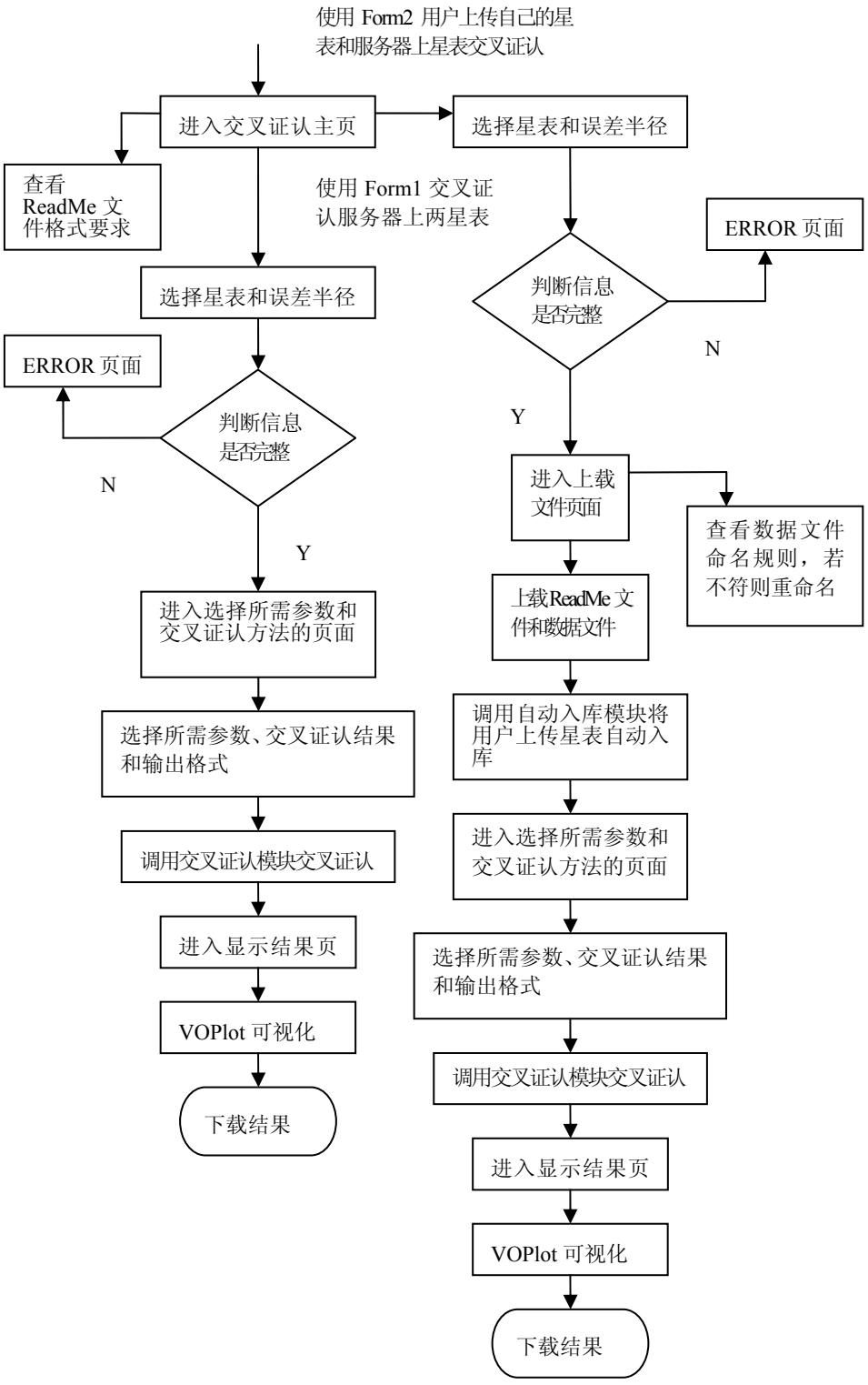


图 3-11 交叉证认工具程序流程图

这个交叉证认工具的核心是交叉证认模块和自动入库模块，它们实现了主要的自动入库和交叉证认功能，这两个模块的具体细节将在下两小节详细说明。模块使用 JavaBean 组件来实现，页面则使用 JSP 来实现，而后端数据库用的是 MySQL，这样就实现了程序和页面的分离。

图 3-10页面提交后下载的结果是一个.dat 文件，如图 3-12所示。图 3-10中参数选择了第一个表 1rxs 的 Count 和第二个表 trc_1 的 BT、VT，而交叉证认结果选择了一对一。此文件共有 8 列，从左至右分别是表 1rxs 的赤经、赤纬、表 1rxs 的 Count 参数、表 trc_1 的赤经、赤纬、表 trc_1 的 BT、VT 参数、角距离 d，其中角距离 d 的单位是角秒。图 3-13是用 VOPlot 工具可视化交叉证认的结果。

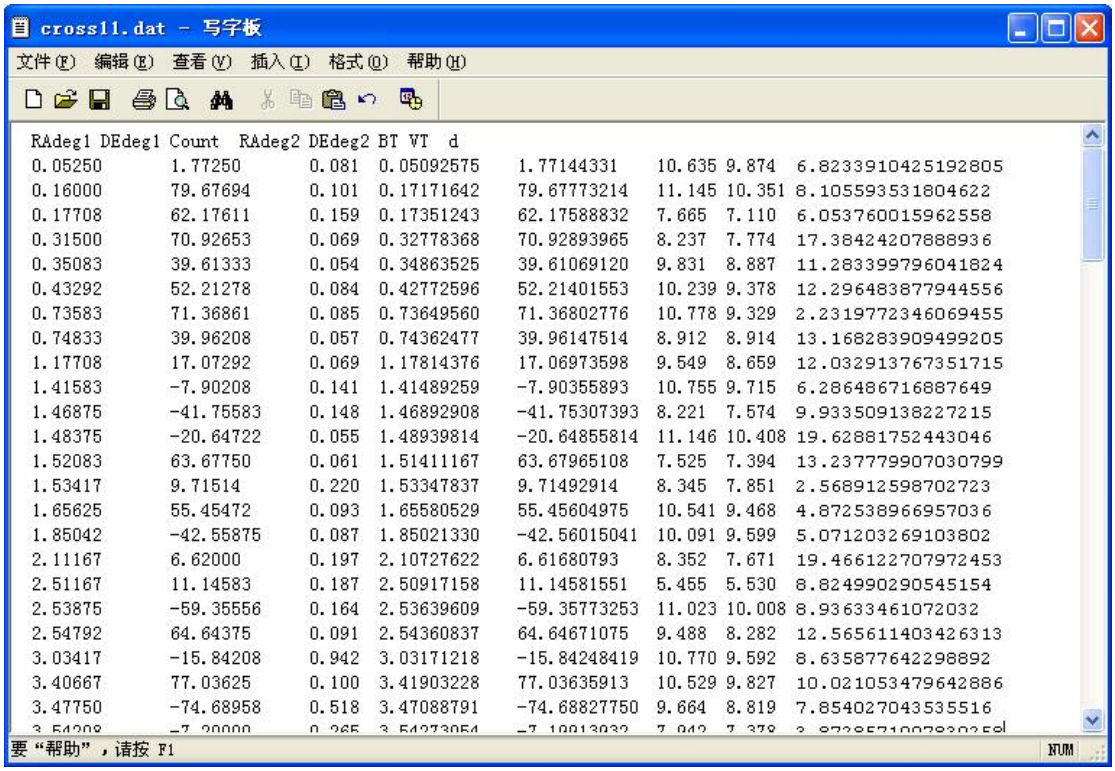


图 3-12 一对一的结果

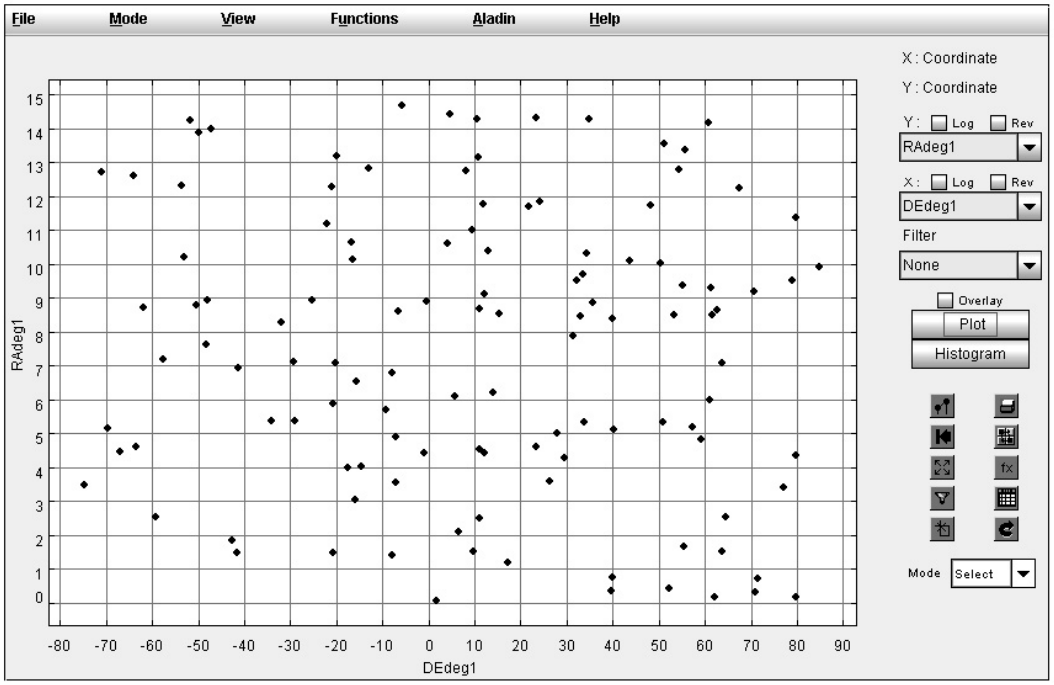


图 3-13 VOPlot 可视化结果

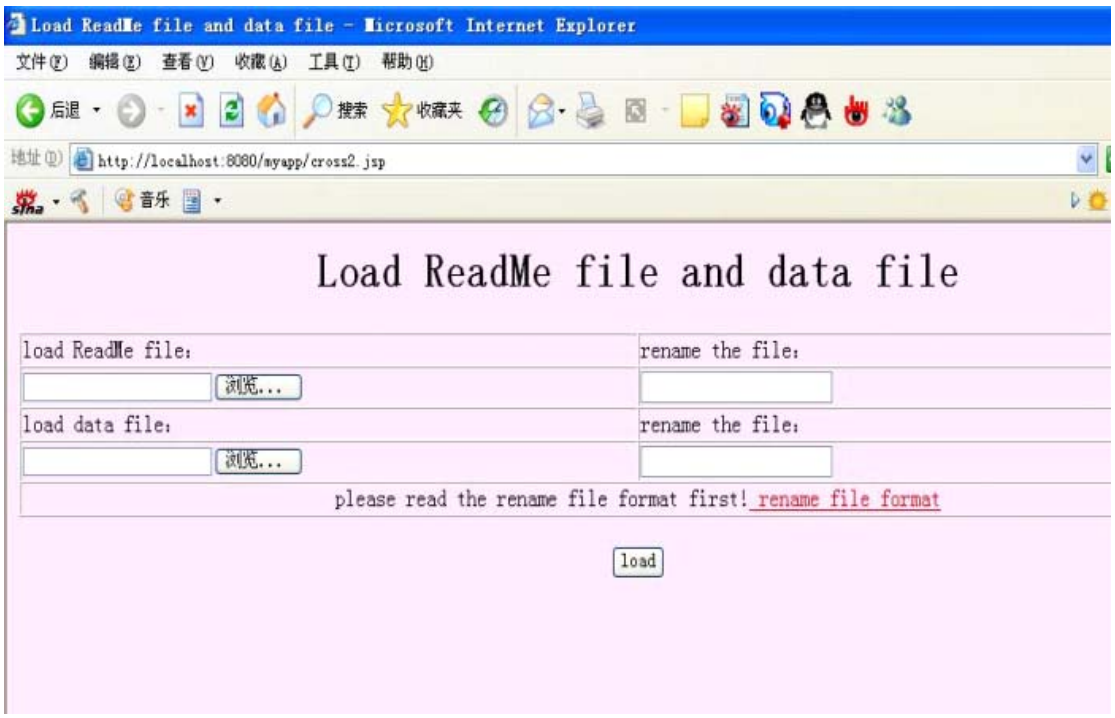


图 3-14 用户上传自己星表的上载文件页面

3.3.4 自动入库模块

自动入库模块不仅实现了根据星表 ReadMe 文件自动入库，也实现了数据库星表自动预处理。输入接口只有数据文件和星表 ReadMe 文件的路径。只要用户上传这两个文件，模块就会自动读取星表 ReadMe 文件并将数据表自动录入数据库，完全屏蔽掉了星表 ReadMe 文件及数据表本身的细节。参照星表 ReadMe 文件标准，用户亦可自己编写或修改星表 ReadMe 文件。这样已经进行了预处理的入库的星表就可直接被交叉证认模块所调用。

自动入库模块接口：

public void setFilepath(String filepath){…}　　将数据文件路径输入模块

public void setReadMefilepath(String ReadMefilepath) {…}　　将星表 ReadMe 文件路径输入模块

public String getMessage(){…}　实现自动入库功能，返回状态值到页面（状态值共有五种， IO error：文件 IO 错误；Database error：数据库错误；ReadMe file Error：ReadMe 文件不符合格式要求；data file Error：上传入库的数据文件不是该 ReadMe 文件中描述的文件；success：自动入库成功）。

3.3.5 交叉证认模块

天文星表的数据量大，且交叉证认的时间复杂度为 O（N^2），这些交叉证认工作的技术难点带来了以下三个问题。首先，内存不够用，为了使工作能够进行，必须设法节约内存，为此我们采用先取部分参数交叉证认，再根据 id 把其他需要参数取出的办法；其次，耗时长，必须设法减少计算量，提高速度，因此我们先确定两星表赤纬覆盖相同天区范围，再画框缩小范围，最后判断对应体，这样可以尽量早地做排除以优化交叉证认速度；最后，数据库查询计算性能低，尤其对于大数据量的查询计算，因此我们对大星表按赤纬分区交叉证认，把任务分解，前面做的数据库星表预处理工作也是为了提高数据库性能。

交叉证认模块算法解决方案：

(1) 以位置精度高的表为中心；

(2) 确定两星表 A、B 的赤纬覆盖相同天区范围：

DEC≤min(max(DEC_A),max(DEC_B))，

DEC≥max(min(DEC_A),min(DEC_B))。

(3) 对大星表按赤纬分区交叉证认；

(4) 先取两星表赤经、赤纬、误差半径、id 参数交叉证认；

(5) 根据 id 把其他需要参数取出；

(6) 画框 ：（在框的范围内才可能是对应体）

$$|RA\_A - RA\_B| < \frac{|r1| + |r2|}{\cos((DEC\_A + DEC\_B)/2)}$$

$$|DEC\_A - DEC\_B| < |r1| + |r2|$$

(7) 判断对应体 ：

$$d \le 3\sqrt{r_1^2 + r_2^2}$$

(8) 对交叉证认的结果分类（一对一、一对多、一对无、无对一）。

交叉证认模块接口：

public void setArgs(String args) {…} 用一个字符串将交叉证认所需信息输入模块，字符串中信息包括：表名、误差半径、所选参数、所选交叉证认结果、结果存储路径等。

public String getMessage() {…} 实现交叉证认功能，返回状态值到页面。状态值共有三种， IO error：文件 IO 错误；Database error：数据库错误；success：交叉证认成功。

3.4 海量星表数据融合系统（XMaS_VO）的实现

为了解决目前存在的数据服务的问题，满足对海量数据处理和交叉证认服务的需求，我们研究并实现了一个不依赖于特定数据库的海量星表数据融合系统（XMatch Sytem of Virtual Observatory，简称为 XMaS_VO），用户能够方便地使用此系统的服务进行星表的上传及自动入库、交叉证认等工作。此外，XMaS_VO 具备可移植性，如果用户需要进行很大数据量的或者大型星表的交叉证认服务，

可以自己建立数据库，方便地将此工具移植到自己的服务器上，自己使用或者开放服务给他人。移植方法可参考附录 C 中的 XmaS_VO 移植方法。随着该系统的进一步发展和应用，天文学家将可以轻松自如地获得他们需要的多波段融合数据，自由地选取参数和匹配的结果，从而可以更加有效地从事海量数据处理和分析工作。

3.4.1 XMaS_VO 的功能及特点

XMaS_VO 系统基于任何支持 SQL 语言的数据库系统，是面向使用者提供的一项服务。如图 3-15所示，现在此系统架构在北京天文数据中心（BADC）上，BADC 有包括 SDSS、2MASS、USNO 等大型巡天星表。系统已经多次被用于数据挖掘工作的数据处理工作，运行正常。当用户只是需要拿自己的一些小星表进行交叉证认，或者拿自己的小星表与此系统的数据库中的大星表进行交叉证认时，可以上传自己的星表，使用 BADC 的系统提供的服务。此系统也可以方便地移植到任何数据库系统上。当用户有大型星表需要进行交叉证认，或者需要上传大型星表的时候，可能会在网络传输上花费较多的时间，更方便的办法是在本地创建所需服务，直接移植该系统，架构到本地进行使用即可。这样不仅用户可以方便快捷地进行交叉证认，而且可以缓解各大天文数据中心面临的大量存储、传输和计算的压力。图 3-15给出了海量星表融合系统的设计。

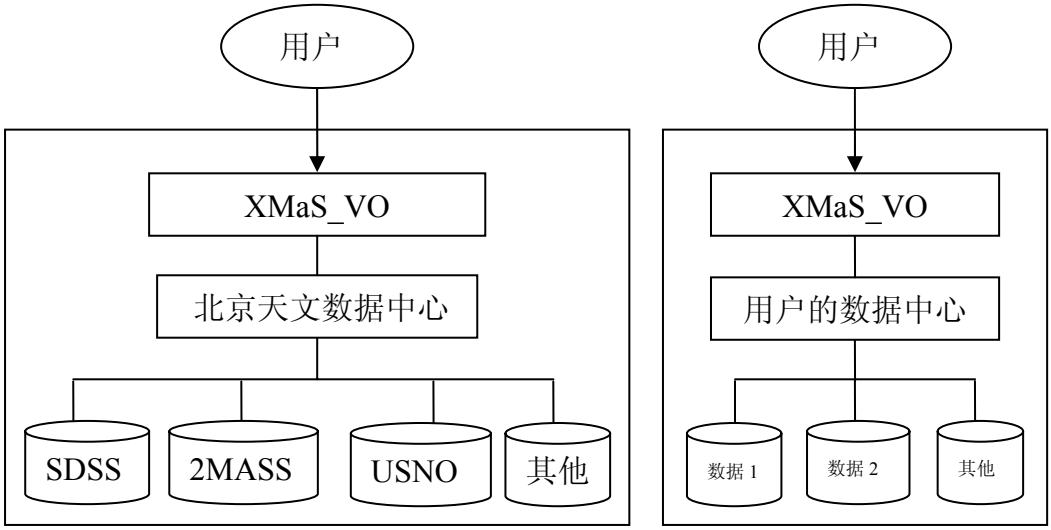


图 3-15 海量星表融合系统的设计

左图：基于服务器的海量星表融合系统；右图：基于用户的海量星表融合系统

XMaS_VO 的功能如下：

（1）可移植性

XMaS_VO 系统不依赖于数据库，可以移植到任何支持 SQL 语言的数据库系统，建立星表数据服务。

（2）突破了数据资源局限

用户可以方便地将需要的数据自动入库到数据库中。

（3）海量数据交叉证认（可以完成两个大巡天项目星表的交叉证认，例如 SDSS 与 2MASS）

通过 HTM 索引分区与 kd-tree 找最近邻算法来实现的。

（4）一对一、一对 N、一对最近邻的交叉证认。

（5）误差半径相近（HTM 级数相同）、相差很大（HTM 级数不同）的交叉证认。

（6）无损耗交叉证认，生成一个新的布尔列，判断是否有对应体。

（7）提取用户需要的参数。

（8）将用户需要的数据导出数据库并传回用户。

3.4.2 XMaS_VO 逻辑过程

XMaS_VO 系统的逻辑过程如图 3-16所示。图 3-16描述了用户使用数据处理系统工作的所有组合情况，用户可以只进行第一步、第二步等工作，也可以是五个步骤地任意组合。为了把五步功能模块整合起来，以更灵活、方便的形式提供给用户，我们采用了批处理技术。图 3-16给出了海量星表融合系统的逻辑过程。

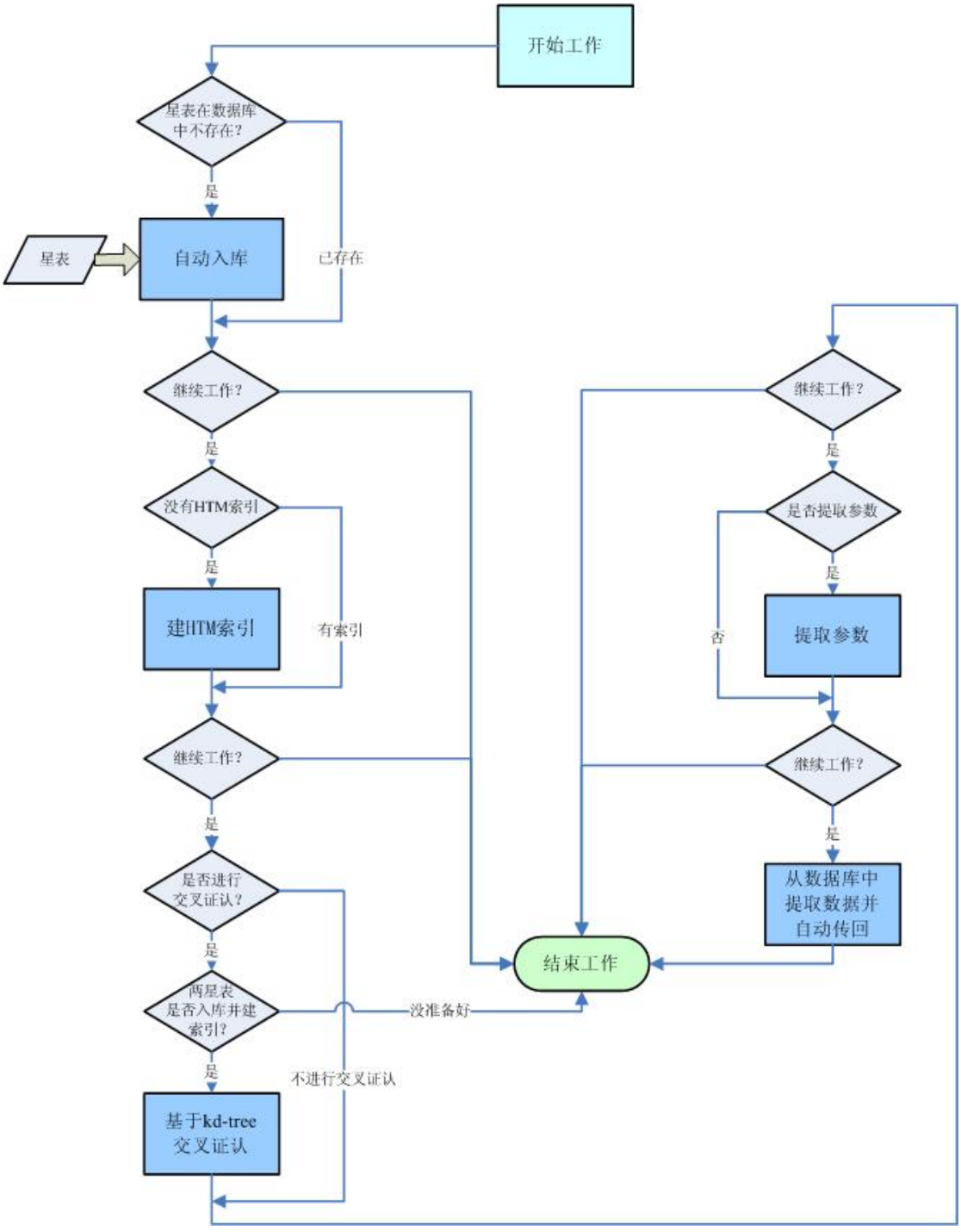


图 3-16 海量星表融合系统的逻辑过程

3.4.3 XMaS_VO 功能模块

XMaS_VO 系统的以上功能由五步功能模块来实现。其中，关键模块是自动

入库、建 HTM 索引以及交叉证认模块。而自动入库模块是将论文第二章中的自动入库工具整合进来的。各功能模块的详细使用说明参见附录 B。

1. 自动入库模块

根据星表 ReadMe 文件，自动建表、入库、将时分秒形式转换成度（RAdeg、DEdeg），建索引，建主键 id_htm。如果用户自己有一批数据，可以参照 ReadMe 文件标准方便地编写 ReadMe 文件。对于 FITS 格式的数据，也可调用本论文第二章提到的程序，将 FITS 文件转换成 ASCII 文件。然后用我们的自动入库程序即可入库。

2. 自动建 HTM 索引模块

根据数据表的坐标数据（单位：度；历元：J2000）计算出对应 HTM 索引的 pcode 值，为星表建立 HTM 索引，将 id_htm 主键和 pcode 值两列新建 HTM 索引表。根据星表误差半径大小确定 HTM 级数的高低。例如误差半径小于等于 5arcsec 时，选 10level；等于 30arcsec 时，选 8level。当两个星表误差半径相差较大时，选择提供两个不一样 HTM 级数的交叉证认程序，可以满足不一样级数的 pcode 值匹配。

3. 交叉证认模块

根据两个数据表及其 HTM 索引表、误差半径进行交叉证认，得到交叉证认表，并分别给交叉证认表中的两表 id 列建索引，方便进一步的数据参数提取。我们为不同的交叉证认需求提供了不同的程序，目前有 10 个，也可以根据交叉证认需求很方便地加入其他程序。

基于 kd-tree 交叉证认，其中对于交叉证认结果，提供了一对一、一对最近邻及一对 N 等；对于误差半径，提供了两星表误差半径和一个交叉证认总体误差等情况；对于 HTM 索引级数，提供级数相同和级数不同两种情况；还能满足无损耗交叉证认的情况，即为中心的星表源的个数没有损耗，交叉证认生成一个新的布尔列，判断是否有对应体。生成的交叉证认表中只有交叉证认两表的赤径、赤纬、id_htm 主键以及角距离 d 的信息，无损耗交叉证认时还有一列布尔列，该列“1”表示有对应体，“0”表示无对应体。

4. 参数提取模块

自动提取交叉证认星表参数，生成交叉证认提取参数表。根据用户需要的参数，对交叉证认表进行操作，生成交叉证认提取参数表。表中包括用户需要的所有参数信息。

5. 数据回传模块

自动从数据库中取出数据并传到本机，指定数据库名、表名等信息，自动从数据库中取出数据，并通过 lftp 传到用户自己的电脑上。

3.4.4 各功能模块的整合

我们在上述各功能模块的基础上，对各个模块完成了进一步的整合工作，使得 XMaS_VO 系统能够批处理地、多任务地进行工作，很大程度地方便了交叉证认等工作的进行。

在进行整合之前，这些功能模块都是独立的程序，每个功能模块使用的时候都需要用户输入很多的参数，很容易出错，易用性方面也不尽人意。而且当需要这些程序共同完成一个工作的时候，例如：先将两个星表入库，然后建立 HTM 索引，再进行交叉证认以及提取参数表，最后再将数据返回给用户，就需要用户分别进行多次参数输入。如星表入库、建立索引就分别需要两次，并且每次工作都必须等待上一步工作完成之后，用户用命令行来执行。这样，不仅需要用户反复查询当前工作是否完成，造成工作的低效，而且用户输入参数本身也是一项繁琐且易出错的工作。

现在各个步骤可以使用参数文件方便的进行编辑，原来需要输入的参数列表，现在可以在功能强大的编辑器中，根据我们提供的模版，编辑这些参数，可以很大地提高用户的效率，且减少了出错的可能性；而将各个步骤的灵活组合，更使得完成复杂功能对于用户而言，只需要一次编辑各步骤的参数文件，无论哪一步需要进行多少次，都只是需要编辑此步骤对应的一个参数文件，然后运行即可。

综上所述，可见整合前和整合后的情况对比如下：

整合前：

（1） 各个工具彼此独立；

（2） 用户需要使用命令行，输入很长的参数表，易出错；

（3） 每次只能执行单个操作；

（4） 易用性差，效率低；

（5） 用户上手难。

整合后：

（1） 使用参数文件使得用户编辑参数更方便；

（2） 提供了参数文件模版，用户更容易理解和使用；

（3） 各个独立的步骤可以灵活组合；

（4） 批处理地连续完成各个功能，不需要用户干预；

（5） 用户学习起来容易。

整合这些工具的基本思路就是使用批处理的思想。整合这些工具使用了 Linux Shell 编程和 vi[90] 的自动处理功能，所以本系统运行于 Linux 操作系统。首先用户根据需要进行的操作编辑批处理文件 steps_crosswork，编辑此文件的时候可以参考我们提供的模板 steps_crosswork_template。还是假设用户需要执行包括将两个星表都入库和建立索引，再将两个星表进行交叉认证，然后进行参数提取和数据传回的步骤。则 steps_crosswork 文件应类似如下：

```
./step1 step1_file_putdata
./step2 step2_file_HTMindex
./step3_7 step3_file_crossmath7
./step4 step4_file_sqldata
./step5 step5_file_getdata
```

这里，step1、step2 等可以被看作命令，具备执行相应的步骤的功能，而 step1_file_putdata 则是 step1 对应的参数文件。这里需要注意，第一步和第二步都事实上对两个星表进行操作，但是这里不需要体现。只需要在相应的参数文件中体现。

举例而言，对 step1 对应的参数文件 step1_file_putdata 可以是：

```
GLMC_l350.tbl ReadMe test glmc
GLMC_l348.tbl ReadMe test glmc
```

这里 GLMC_l350.tbl、GLMC_l348.tbl 就是用户提供的星表，其他几个则是一些参数。

其他各个步骤的设计都与此类似，首先包装原有的命令行产生对应的以待接收参数的文件（即功能文件）。然后构造对应的 vi 脚本文件，用以处理用户提供的参数文件。最后构造 step 文件，统一调用这些脚本。整合各步骤详细说明参见附录 A。

3.4.5 算法实现流程

XMaS_VO 的交叉证认功能模块的算法是采用 HTM 索引分区与 kd-tree 找最近邻算法。此算法是解决海量数据融合问题的有效方案，利用了 HTM 算法进行索引分区，并在每个分区中用 kd-tree 算法寻找最近邻。数据处理系统提供了多个交叉证认功能模块，解决了不同的交叉证认需求，这里简单介绍一下相同 HTM 级数、一对最近邻、两个误差半径、有损耗交叉证认的情况算法实现的流程。

(1) 以小表为中心；

(2) 将小表中的不同 HTM 值取出，生成临时表 A；

(3) 对表 A 中所有 HTM 值循环；

I. 对每个 HTM_i，将大表中 HTM 值为 HTM_i 的源取出 id_htm、RAdeg、DEdeg 列生成临时表 B；

II. 根据表 B 是否为空，判断大表中是否存在 HTM_i 值对应的数据，如为空，跳到步骤(3)的下一次循环；

III. 将小表中 HTM 值为 HTM_i 的源取出 id_htm、RAdeg、DEdeg 列生成临时表 C；

IV. 用表 B 源的 RAdeg、DEdeg 值，建 Kd-tree （insert）

V. 对表 A 中的所有源循环

a. 在 kd-tree 中找最近邻 （nearest）；

b. 测试是否符合交叉证认判断公式 $d \le 3\sqrt{r_1^2 + r_2^2}$；

c. 输出结果。

HTM 索引分区与 kd-tree 找最近邻算法的 SQL 语言描述如下：其中 test.distinct 表是以小表为中心 HTM 分区的表，test.id_ra_decl_big 表是大表某个 HTM 分区中的子集，test.id_ra_decl_small 表是小表在某个 HTM 分区中的子集。这三个表都是由算法程序自动生成并自动删除的中间过程的表，即上面的临时表 A、B 和 C，用户不需要知道这个细节，只需要保证数据库中有 test 数据库即可。

以小表为中心

```
create table test.distinct select distinct pcode from 小表;
（将小表中的不同 pcode 值取出，生成 test.distinct 表）
For each pcode from test.distinct
（对 test.distinct 中所有 pcode 值循环）
{
    create table test.id_ra_decl_big select id_htm，RAdeg，
Dedeg from 大表 where pcode=pcode_i;
    （将大表中 pcode 值为 pcode_i 的取出列 id_htm、RAdeg、Dedeg）
    if test.id_ra_decl_big != null
  {
     create table test.id_ra_decl_small select id_htm，RAdeg，
Dedeg from 小表 where pcode=pcode_i;
    （将小表中 pcode 值为 pcode_i 的取出列 id_htm，RAdeg，Dedeg）
     给表 test.id_ra_decl_big 建 kd-tree （insert）
     For each RAdeg,Dedeg from test.id_ra_decl_small
```

```
    {
        在 kd-tree 中找最近邻  （nearest）或 N 个最近邻
        测试是否符合 d ≤ 3√(r1^2 + r2^2)
        输出结果
    }
  }
}
```

3.5 交叉证认在 VO-DAS 中的应用

3.5.1 VO-DAS 简介

VO-DAS[32] 是一个基于网格技术的虚拟天文台数据访问系统，支持对异地异构海量数据资源的统一、无透明访问。

统一访问这些分布存放的异构数据这个课题，一直是各国虚拟天文台的研究重点。NVO、JVO、CDS 等都进行了很多尝试。到目前为止，虽然已经有若干数据访问方案提出，还没有一种能完全满足天文学家的需求的服务，它们要么因为所选择的网络技术而限制了其功能，要么因为为专门的项目服务而限制了应用范围和互操作性。一个理想的天文数据访问系统应该具备以下特点：

- 能够访问分布保存的异构天文数据库；
- 支持自动发现数据资源；
- 支持特定空间区域的数据访问；
- 支持海量数据访问；
- 支持天文数据的交叉证认；
- 支持星表数据、图像数据和光谱数据的访问；
- 具有标准的数据访问接口；
- 具有良好的互操作性，可以和任何符合 IVOA 规范的 VO 应用程序互联。

依据这样的需求，我们 VO 小组设计了中国虚拟天文台自己的数据访问服务（VO-DAS）。此系统建立在 Globus Toolkits 4（GT4）[33] 之上。它逐步能够实现异地异构数据库的访问、数据资源的自动发现、海量数据的访问与查询、多种数据格式的访问、天文数据的交叉认证等。此外，它的对外接口简单实用，可以针对不同需求的天文数据用户发展出多种网格应用服务。

3.5.2 VO-DAS 中的交叉证认

如图 3-17所示，客户端通过 VO-DAS 对多个数据资源进行带有交叉证认功能的查询步骤如下：

1) Client 提交 ADQL（天文数据查询语言）语句。
2) VO-DAS 解析 ADQL，生成 ExecPlan（此组件的详细设计可参见附录 D）。ADQL 中涉及几个 DataNode，则在 ExecPlan 生成几个 SQL 语句。
3) ADQL 语句中有用于交叉证认功能的 XMATCH 函数，需要在 ExecPlan 中把此函数替换成相应的 SQL 语句。
4) 将任务分别提交到相应的 DataNode 并执行。
5) 对于交叉证认功能，需要首先在第一个 DataNode 上进行查询；将查询结果传输到第二个 DataNode 上并为此查询结果建立临时数据表，把此数据表和第二个 DataNode 上的查询结果进行交叉证认；依此方法扩展到最后一个 DataNode，在其上进行交叉证认并将结果存放在 MySpace 上。

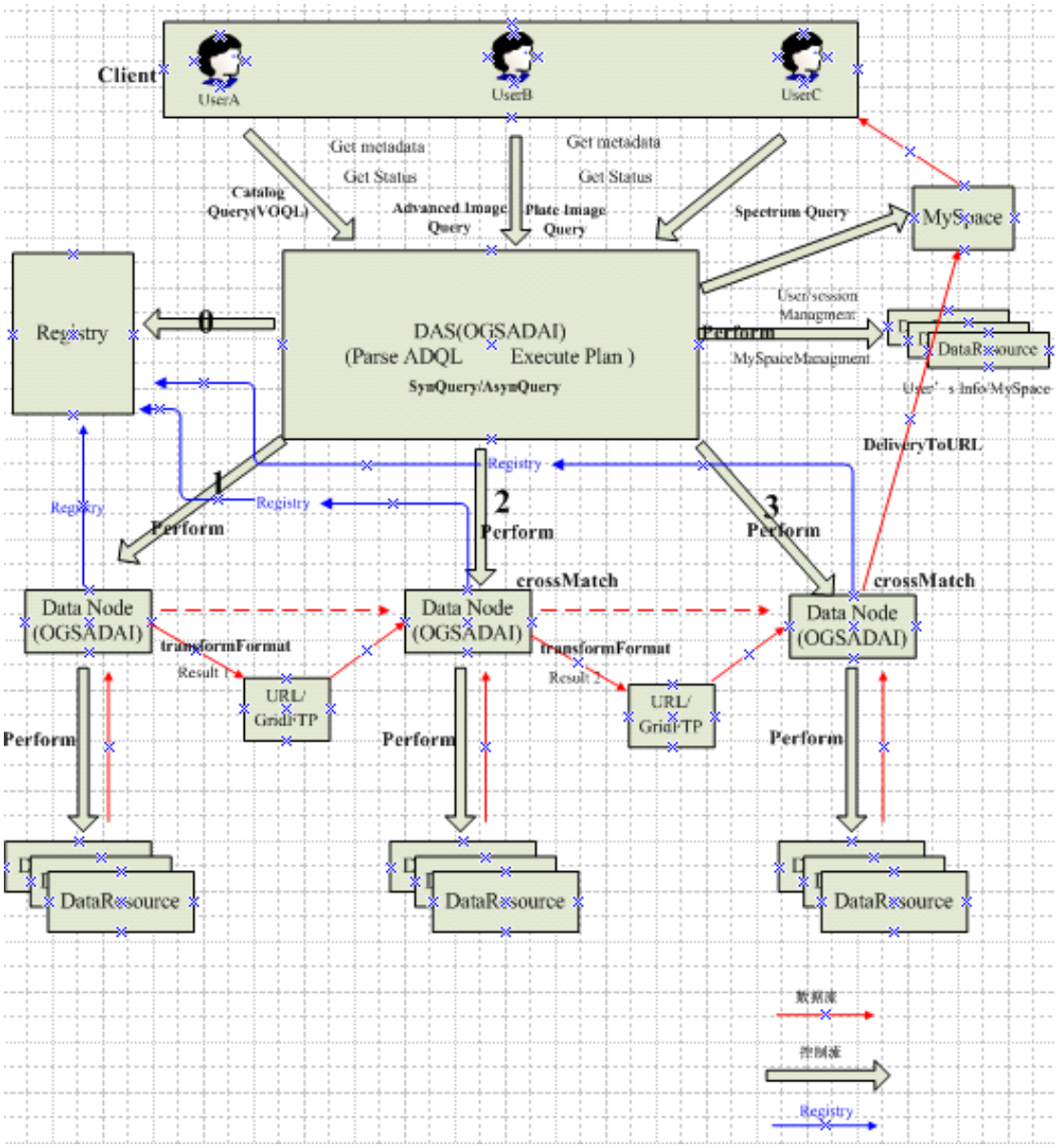


图 3-17 带有交叉证认功能的查询流程图

ADQL 是根据 SQL 语句改进并针对天文数据访问特点而设计的数据库查询语言，它继承了绝大多数 SQL 语言的语法特征并带有一定的扩展性。

VO-DAS 将 ADQL 拆分成 SQL 语句的方法举例如下：

1) 客户端用来查询的 ADQL 语句如下：

```
SELECT o.objId, o.ra,
    o.dec, o.r, o.type,
```

```
    t.objId, t.ra, t.dec
FROM
    SDSS:PhotoPrimary o, TWOMASS:PhotoPrimary t
WHERE XMATCH(o, t) < 3.5
```

2) 上面的语句涉及到两个 DataNode，所以将它拆分成两个 SQL 语句：

```
SELECT o.objId, o.ra,
    o.dec, o.r, o.type
FROM
    SDSS:PhotoPrimary o
```

以及

```
SELECT up.objId, up.ra,
    up.dec, up.r, up.type,
    t.objId, t.ra, t.dec
FROM
    Upload up, TWOMASS:PhotoPrimary t
WHERE XMATCH(up, t) < 3.5
```

3) 将 XMATCH 函数替换成为相应的 SQL 语句：

```
SELECT up.objId, up.ra,
    up.dec, up.r, up.type,
    t.objId, t.ra, t.dec
FROM
    Upload up, TWOMASS:PhotoPrimary t
WHERE
    3.5*3.5/1200/1200 >
   ((up.ra_up-t.ra)*cos((up.decl_up+t.decl)*3.14/180/2))^2
    +(up.decl_up-t.decl)^2
```

3.6 测试结果与应用现状

3.6.1 模拟数据检验交叉证认准确率

为了验证 HTM 索引分区与 kd-tree 找最近邻算法和 XMaS_VO 系统进行交叉证认的准确率，我们设计了模拟数据来测试。模拟数据可以用任何星表数据作为原始数据，以及被人为在原始数据的相应列（例如赤经、赤纬和星等）中加入噪声的数据。把这两个数据进行交叉证认，就可以验证本系统的准确率。

模拟数据的生成：原始数据是 GSC2.3。对 6.5×6.5 平方度的施密特底片数

据进行编号，并根据实际星表间系统误差和随机误差的水平将噪声加到星表数据中，从而再生成另一张底片数据。这样一方面可以确保模拟的真实性，同时也可以在证认前就知道每颗天体在两个星表数据间的对应关系，以便检验证认方法的可行性和准确率。

模拟数据包括两个文件：原始数据文件和加入位置与星等噪声的数据文件。为了尽可能接近真实的情况，根据目标在底片上的不同位置，加误差的量级和形式是不一样的，数据的随机误差量级是 0.2 角秒，系统误差接近 0.0 角秒，具体请见图 3-18。

数据共 6 列：第 1 列为目标的代码，第 2 列为赤经（度），第 3 列为赤纬（度），第 4 列为星等（mag），第 5 列为底片坐标 X，第 6 列为底片坐标 Y。同一颗天体在两个文件中的代码相同，只有赤经、赤纬、星等不同，用交叉证认工具找好对应体后，只要比较一下目标代码是否一样，就可以知道是否找对。

模拟数据文件样例：

1 206.0597425 38.3316676 12.9000 3677.42 823.14

2 206.0148827 38.3227335 14.4100 3751.54 802.64

3 206.1054124 38.3151916 13.3500 3600.79 789.93

4 206.2188895 38.3074940 14.9100 3411.86 777.95

5 205.9300483 38.3085064 14.8000 3891.90 769.61

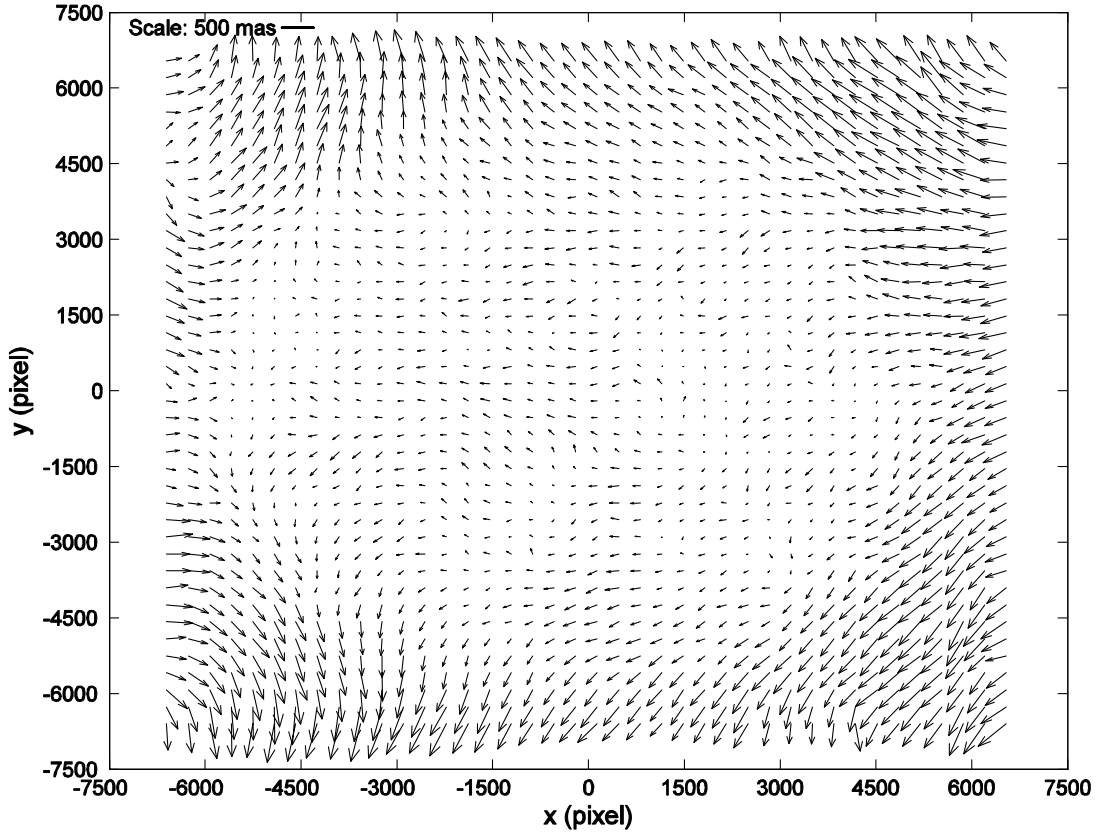


图 3-18 模拟数据噪音

模拟数据共 295,832 行，交叉证认结果如下：

1. 选择 HTM Level=10，耗费时间约 300 多秒。

当误差半径 r1=r2=0.2，共有 282282 (95.42%) 交叉证认结果，其中正确的 282176(106 个不正确)，占 99.96%

当误差半径 r1=r2=0.25，共有 292069 (98.73%) 交叉证认结果，其中正确的 291960(109 个不正确)

当误差半径 r1=r2=0.4，共有 295288 (99.82%) 交叉证认结果，其中正确的 295179(109 个不正确)

2. 选择 HTM Level=5，耗费时间约 2000 多秒。

当误差半径 r1=r2=0.2，共有 282746 (95.58%) 交叉证认结果，其中正确的

282640(106 个不正确)，占 99.96%

当误差半径 r1=r2=0.25，共有 292570 (98.90%) 交叉证认结果，其中正确的 292461(109 个不正确)

当误差半径 r1=r2=0.4，共有 295806 (99.99%) 交叉证认结果，其中正确的 295697(109 个不正确)

交叉证认测试选择了两种 HTM Level。其中 Level=10 是按照误差半径选择的，小于 5 角秒的都选 Level=10。当 Level=5 时分区较大，误差半径为 0.4，所有数据基本都证认上了（99.99%），可见我们的交叉证认算法的结果是可靠的。结果显示 Level=5 时耗费时间是 Level=10 的近十倍，可见用 HTM 索引分区能提高交叉证认的效率。但 Level=10 时有 500 个左右的数据证认不上，可能因为对应体被分到不同的 HTM 分区上，可见分区会带来小量的证认错误。所有结果都约有 109 个数据证认结果不正确，可能这些数据的误差比较大，另外又有比较近的源，所以就误判了。结果显示采用了 HTM 索引分区与 kd-tree 找最近邻算法的 XMaS_VO 系统是可靠、高效、灵活的。

3.6.2 XMaS_VO 的应用

目前 XMaS_VO 系统已经成功应用于很多海量星表的交叉证认工作，如表 3-2所示。一般两个几十万或几百万条数据的大星表交叉证认大概需要一个小时左右，而一个几百万条数据的大星表和 2MASS、USNO-B1.0 这样几亿或几十亿条数据的海量星表交叉证认需要十几小时到二十几小时的时间。可见通过建立 HTM 索引分区并在每个分区内用 kd-tree 算法，可以实现在可接受的时间内完成海量星表的交叉证认功能。

表 3-2 XMaS_VO 应用实例

星表 A	行数	大小	星表 B	行数	大小	HTM 级数	时间
ROSAT2	105,924	18M	Tycho2	2,539,913	439 M	8	3567sec
SDSS qusars	76,989	56M	2MASS	470,992,970	123 G	10	5033sec
FIRST	811,117	83M	2MASS	470,992,970	123 G	10	24404sec
Gspc24	554,007	65M	USNOB	1,045,096,352	172 G	5,8	85720sec
GSC2.3 的部分原始数据	295,832	23M	加入位置和星等噪声的数据	295,832	23M	10	338sec

3.7 小结

本章首先介绍了交叉证认的定义、国内外研究情况以及技术难点分析等内容。为了解决内存限制和交叉证认算法复杂度高的问题，研究了两种交叉证认算法：B-tree 索引算法和 HTM 索引算法。在此基础上，创新性地提出了 HTM 索引分区与 kd-tree 找最近邻算法，它借鉴了前两种算法的优点，并引入了数据挖掘领域 kd-tree 算法找最近邻的思想，彻底解决了内存和效率的瓶颈，具有较高的证认精度。接着，本章介绍了交叉证认工具的开发、天文星表融合系统（简称 XMaS_VO）的开发以及交叉证认工具在 VO-DAS 中的应用。XMaS_VO 系统是利用 HTM 索引分区与 kd-tree 找最近邻算法进行数据融合的，它整合了自动入库工具、HTM 索引和交叉证认工具。最后用模拟数据对 XMaS_VO 系统做交叉证认的准确率和效率进行了测试，结果显示 HTM 索引分区与 kd-tree 找最近邻算法运用于 XMaS_VO 系统中是可靠、高效、灵活的。XMaS_VO 已经用于很多海量天文数据融合的应用实例中。

第 4 章 天文数据挖掘研究

数据挖掘技术可以有效地应用于天文学中，用以发现新类型的天体，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，都将遇到自动分类、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。

有了强大的天文数据融合系统 XMaS_VO，大大减少了数据挖掘的数据准备和预处理的工作量，从而使海量天文数据挖掘课题研究成为可能。这样算法研究人员只需关注在算法性能和应用的研究本身，而无需对海量数据处理投入太多的时间和精力。

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、回归或聚类等研究。

天文学从始至终都包含着分类和回归问题，传统的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，而且当天文数据的信噪比低时，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，如像 SDSS 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和星系测光红移评估是不可能的。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，以快速准确地完成天体的自动分类和回归，而这一切必须依靠强有力的分析技术支持才能完成。

4.1 分类和回归研究

4.1.1 分类研究

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，不能完全确定。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。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。

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。因此，通常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[34] [35]，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[36] [37] [38]，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）[39] [54]，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。

具体的天文应用如：Hatziminaoglou 等人研究了一种避免通常偏差的新方法来通过测光数据区分类星体和恒星/星系[40]。Wolf 等人研究了一种测光方法在多色巡天中识别恒星、星系和类星体[41]。McGlynn 等人使用决策树建造了一个在线系统来自动分类 X 射线源[42]。Carballo 等人用神经网络算法从射线和光学波段挑选出类星体候选体[43]。Suchkov 等人开发了一种倾斜的决策树分类器 ClassX 来优化基于 SDSS 测光星表数据的天体分类和测光红移评估[44]。Ball 等人使用决策树对 SDSS DR3（Data Release 3）的恒星和星系数据分类[45]。神经网络算法被用于星系的光谱分类（Storrie-Lombardi 等人[46]、Adams 和 Woolley 等人[47]）。学习矢量量化（LVQ）被用于天体分类[48]。贝叶斯信任网络（BBN）、多层感知器（MLP）网络和交互式决策树（ADtree）被用来比较分类类星体和恒星的效率[49]。决策树，比如 REPTree、Random Tree、Decision Stump、Random Forest、J48、NBTree 和 ADTree 被用来研究分类活动天体和非活动天体[50]。

4.1.2 回归研究

回归分析是以客观事物变量间的统计关系为重要研究对象，基于对客观事物进行大量实验和观察，寻找隐藏在不确定的现象中的统计规律的统计方法。统计回归问题就是从已知的资料出发，建立自变量 x 对因变量 y 的回归方程，以便利用这个回归方程进行新的预测。

分类和回归都可用于预测，两者的目的都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。与回归不同的是，分类的输出是离散的类别值，而回归的输出是连续数值。二者常表现为决策树的形式，根据数据值从树根开始搜索，沿着数据满足的分支往上走，走到树叶就能确定类别。

根据因变量 y 和自变量 x 之间是否存在线性关系，回归问题可分为线性回归和非线性回归。根据给定资料数据的已知条件不同，回归问题可分为参数回归和非参数回归。如果数据的分布是不确定的，自变量和因变量之间关系未知，称为非参数回归。非参数回归首先要对数据的分布进行估计，然后才能确定回归函数。

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。

天文中的测光红移预测问题归根结底属于数据挖掘的回归问题。需要用回归的算法来实现。目前已有许多这方面的工作，有模版匹配、最近邻、神经网络、高斯过程、多变量多项式回归、支持矢量机、核回归、复合算法等，探讨不同的参数选择与不同的回归方法对红移测量精度的影响。1957 年 Baum 提出利用测光数据研究红移[51] 。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[52] 。这种方法与光谱红移相比有很大的优势。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。测光红移最常用的方法是模板匹配，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[53] 。另一种方法是使用多项式或其他函数进行拟合，Connolly[54] 和 Sowards-Emmerd[55] 分别作了将红移拟合为星系颜色多项式的工作。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。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，简称 SVMs）[56] ，神经网络[57] [58] [59] 等。

4.2 算法原理

本章采用监督的 kd-tree、支持矢量机（SVMs）和随机森林（Random Forest）算法对天文数据进行分类和回归，目的是将类星体从恒星或恒星和星系的样本中分离出来，以及从测光数据中预测星系的测光红移。

4.2.1 kd-tree 算法

kd-tree（k-dimensional tree）作为一个计算机科学术语，指的是一个用于在 k 维空间中组织数据分布的空间分区数据结构[60] ，其中 k 代表空间维数。kd-tree 使用的是于坐标轴垂直的分割平面。kd-tree 从树根到叶子都存储一个数据点，所以每个分割平面都必须穿过 kd-tree 中的一个数据点。kd-tree 在 k 维空间内组织数据点集合的方式如下：其一旦构造好，当收到列出邻接的所有数据点的请求时，无须扫描每个数据点就可以快速地响应。每一个树节点表达参数空间的一个子集，其根节点包含全部 k 维参数空间。无叶子的节点有两个孩子。通过将父节点的边界框的最广的维度进行划分，将在此维度上比分割点的数值小的点数据分配给左孩子，其余的分配给右孩子。kd-tree 常常自顶而下地构造，开始于全部点的集合，然后在最广的维度的中心进行分割。这样就产生了两个子节点，每个节点拥有不同的数据点。通过递归地进行上述步骤可构造一个完整的 kd-tree。

当 k=3 时，3 维 kd-tree 也可称为 3D tree，如图 4-1所示，第一次分割将根单元分成两个子单元，它们进一步被分割成为两个子单元。然后这四个中的每一个继续被分割成两个子单元。直到无法进行更多的分割，最终的八个被称为叶子单元。灰色的球面代表树顶。

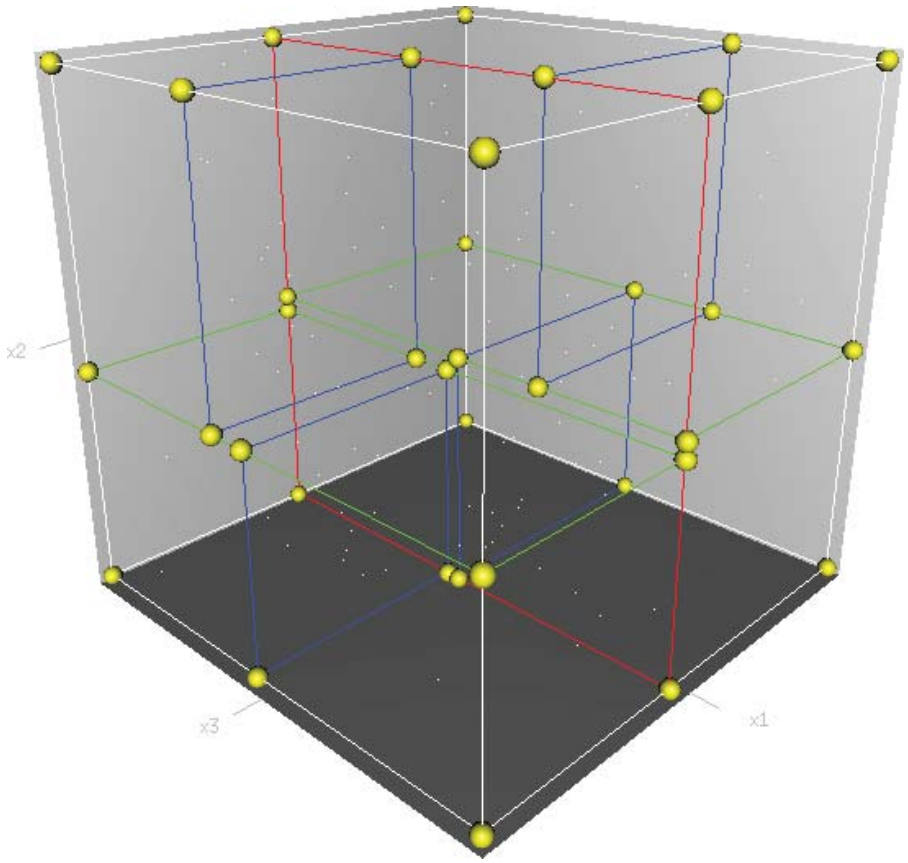


图 4-1 三维 kd-tree

kd-tree 算法被用于划分大型的 SDSS 科学数据，以 5 个星等作为参数建立索引[61] 。Maneewongvatana 和 Mount 提出了两种新的 kd-tree 划分方法的实验性分析[62] 。Hsieh、Yee 和 Lin 等人使用 kd-tree 算法来划分他们的样本来提高星系红移的准确率[63] 。Kubica 等人使用 kd-tree 提高小行星探测的效率[64] 。

4.2.2 SVMs 算法

SVMs（Support Vect Machines）是一种用于分类和回归的监督统计学习方法，其理论由 Vapnik 于 1995 年提出[65]，它可视为 Tikhonov 正规化的一种特例。它将输入向量非线性地映射到高维特征空间并在此空间构造最优分割超平面。SVMs 具有很多变量，其中核函数用于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，而不是在高维特征空间中进行。核函数有多种形式，比较流行的包括线性核函数、多项式核函数、径向积核函数等。

SVMs 的主要思想可以概括为两点：(1) 它是针对线性可分情况进行分析，对于线性不可分的情况，通过使用非线性映射算法将低维输入空间线性不可分的样本转化到高维特征空间使其线性可分，从而使得高维特征空间采用线性算法对样本的非线性特征进行线性分析成为可能；(2) 它基于结构风险最小化理论之上在特征空间中建构最优分割超平面，使得学习器得到全局最优化，并且在整个样本空间的期望风险以某个概率满足一定上界。支持向量机的目标就是要根据结构风险最小化原理，构造一个目标函数将两类模式尽可能地区分开来，通常分为线性可分和线性不可分两类情况。

SVMs 算法已经在天文学中有许多成功应用，比如：变星分类[66] [67] [68]，星系形态分类（Humphreys 等 2001），太阳耀斑探测[69]，多波段数据分类[70] [71]，星系测光红移评估[56] 和天体物理学的不同天体星表的证认[74] [75]。王丹等人[72] [73] 研究了 SVMs 和核回归算法用 SDSS DR5（Data Release 5）和 2MASS 数据进行测光红移的评估。

4.2.3 随机森林算法

随机森林(Random Forest)是一种利用多个分类树对数据进行判别与分类的方法。它可以应用于变量个数远大于样本个数的数据，并且不会产生过拟合现象，在对数据进行分类的同时，可以给出各个参数在分类过程中的重要性评分，根据该评分能够筛选出相对重要的变量。

随机森林由 Leo Breiman 在 2001 年提出[76]，它通过自助法(bootstrap)重采

样技术，从原始训练样本集中有放回地重复随机抽取个样本生成新的训练样本集合，然后根据自助样本集生成多个分类树组成随机森林，新数据的分类结果按分类树投票多少形成的数而定。

随机森林中的每个分类树使用下面的算法生成：

(1) 原始训练集为，应用 bootstrap 法有放回地随机抽取个新的自助样本集，并由此构建棵分类树，每次未被抽到的样本组成了袋外数据(out-of-bag，OOB)。

(2) 对每个变量，则在每一棵树的每个节点处随机抽取该变量，然后根据节点不纯度度量的 Gini 准则，在其中选择一个最具有分类能力的变量，变量分类的阈值通过检查每一个分类点确定。

(3) 每棵树最大限度地生长，不做任何修剪。

(4) 将生成的多棵分类树组成随机森林，用随机森林分类器对新的数据进行判别与分类，分类结果按树分类器的投票多少而定。

随机森林（RF）算法已经在天文学中得到广泛应用。比如，Breiman 等人用这个方法来证认 FIRST 巡天中的类星体[77] 。Albert 等人应用于基于地面的 γ 波段望远镜 MAGIC 的数据分析[78] 。Carliles 等人用 SDSS DR6 数据的颜色参数和光谱红移来进行星系红移评估[79] 。Bailey 等人给出了寻找超新星的分类结果[80] 。

4.3 用 kd-tree 和 SVMs 算法分类类星体和恒星

基于可见光（SDSS DR5）和近红外（2MASS）波段的巡天数据，应用 kd-tree 和 SVMs 算法来将类星体与恒星区分开。考察不同的参数组合对 kd-tree 和 SVMs 算法的影响，并且用不同的性能测试标准来选择分类器模型。

4.3.1 巡天

SDSS（the Sloan Digital Sky Survey）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[81] [82] 。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。

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，最终得到的每根光谱覆盖的波长范围为 3800 埃到 9200 埃，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。

SDSS 的 DR5（Data Release 5）数据包括 2005 年 6 月前获得的全天高质量巡天数据，并标志着 SDSS 巡天第一期工程的结束。它包括围绕北银冠 8000 平方度 215 百万个天体的 5 个波段的测光数据，和从 5740 平方度的光谱图像数据中提取出来的 1,048,960 个星系、类星体和恒星的光谱数据。

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[83] 。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。对于点源星表：470 百万个源在 JHKs 三波段精确的位置和测光方面的信息，大部分是银河系内的恒星，也包括一些不可分辨的源。对于展源星表：1 百万多个源在 JHKs 三波段的位置和累计星等方面的信息，97%是星系，也包括银河系内一些可分辨的源。

4.3.2 样本

我们收集了 SDSS DR5 有光谱的类星体和恒星的测光数据，并通过海量星表融合系统 XmaS_VO 与 2MASS 星表交叉证认，选择 2 角秒的误差半径。表 4-1 中描述了我们获得的样本数据，结果显示几乎所有的 SDSS 源在 2MASS 星表中都有对应体。

表 4-1 星表样本数量

星表	类星体数量	恒星数量
SDSS	76,949	108,744
SDSS+2MASS	76,863	108,679

为了研究类星体和恒星在多维空间的分布，我们使用了 SDSS 数据的不同星等：包括 PSF 星等（u^p g^p r^p i^p z^p）、模型星等（u g r i z）和经过红化校正的模型星等（u' g' r' i' z'），以及 2MASS 星表的 J、H 和 K_s 星等，我们用这些参数的组合来探索不同的输入参数对分类结果的影响。表 4-2中总结了样本参数的统计特性，第一、二和三列分别给出了参数号、参数名和参数描述，其他列给出了类星体和恒星数据中参数的均值和方差。很明显，不同类型的源参数的统计特性不同，尤其是颜色参数。因此，用这些参数来分类类星体和恒星是合理和适用的。为了研究不同天体在二维散点图上的分布，我们挑选了一组参数（u-g、g-r、r-i、i-z、r）对类星体和恒星数据抽样样本画散点图，如图 4-2所示。另外，我们用图 4-2中同样的参数，将主分量分析方法（PCA）应用在样本上。PCA 是一个用于在较大数目的观测属性中判断出最小数目的独立或非相关的变量的方法。PCA 算法可用于数据压缩与分析以及无监督分类。PCA 算法在天文中的应用很多，比如 Connolly 和 Szalay 等人[39] 和 Zhang 和 Zhao[70] 的研究。PCA 算法的结果显示前三个特征向量所占的百分比分别为 99.30%、0.41%和 0.17%，这表明前三个向量携带了几乎所有的信息，尤其是第一个向量。我们研究主分量空间中类星体和恒星的分布。其中，PC1、PC2 和 PC3 分别是第一、二和三主分量的简称。从图 4-2中很明显看出，用双色图或主分量空间很难将类星体和恒星区分开。因此我们需要基于机器学习或数据挖掘技术来实现类星体和恒星在高维空间中的分类。

No.	Parameters	Desciption	Quasars	Stars
1	u^p	SDSS PSF u magnitude	19.78 ± 1.38	20.78 ± 2.57
2	g^p	SDSS PSF g magnitude	19.31 ± 0.89	19.32 ± 2.23
3	r^p	SDSS PSF r magnitude	19.06 ± 0.76	18.56 ± 1.81
4	i^p	SDSS PSF i magnitude	18.89 ± 0.75	18.02 ± 1.51
5	z^p	SDSS PSF z magnitude	18.80 ± 0.76	17.73 ± 1.48
6	$u^p - g^p$	SDSS PSF $u-g$ color	0.48 ± 0.79	1.46 ± 1.04
7	$g^p - r^p$	SDSS PSF $g-r$ color	0.25 ± 0.35	0.76 ± 0.81
8	$r^p - i^p$	SDSS PSF $r-i$ color	0.16 ± 0.20	0.54 ± 0.87
9	$i^p - z^p$	SDSS PSF $i-z$ color	0.10 ± 0.18	0.30 ± 0.59
10.....	u	SDSS model u magnitude	19.74 ± 1.41	20.77 ± 2.63
11.....	g	SDSS model g magnitude	19.23 ± 0.93	19.26 ± 2.24
12.....	r	SDSS model r magnitude	18.96 ± 0.84	18.49 ± 1.83
13.....	i	SDSS model i magnitude	18.81 ± 0.85	17.96 ± 1.52
14.....	z	SDSS model z magnitude	18.71 ± 0.87	17.67 ± 1.50
15.....	$u-g$	SDSS model $u-g$ color	0.50 ± 0.81	1.50 ± 1.09
16.....	$g-r$	SDSS model $g-r$ color	0.26 ± 0.37	0.78 ± 0.85
17.....	$r-i$	SDSS model $r-i$ color	0.17 ± 0.21	0.53 ± 0.89
18.....	$i-z$	SDSS model $i-z$ color	0.10 ± 0.18	0.29 ± 0.62
19.....	u'	SDSS dereddened model u magnitude	19.58 ± 1.41	20.58 ± 2.63
20.....	g'	SDSS dereddened model g magnitude	19.12 ± 0.93	19.13 ± 2.24
21.....	r'	SDSS dereddened model r magnitude	18.89 ± 0.84	18.39 ± 1.83
22.....	i'	SDSS dereddened model i magnitude	18.74 ± 0.85	17.88 ± 1.52
23.....	z'	SDSS dereddened model z magnitude	18.87 ± 0.87	17.61 ± 1.50
24.....	$u'-g'$	SDSS dereddened model $u-g$ color	0.46 ± 0.81	1.45 ± 1.09
25.....	$g'-r'$	SDSS dereddened model $g-r$ color	0.23 ± 0.37	0.74 ± 0.85
26.....	$r'-i'$	SDSS dereddened model $r-i$ color	0.15 ± 0.21	0.51 ± 0.89
27.....	$i'-z'$	SDSS dereddened model $i-g$ color	0.08 ± 0.18	0.27 ± 0.62
28.....	J	2MASS J magnitude	15.58 ± 1.39	15.39 ± 1.21
29.....	H	2MASS H magnitude	15.00 ± 1.33	14.91 ± 1.21
30.....	K_s	2MASS K_s magnitude	14.67 ± 1.23	14.72 ± 1.19
31.....	$J-H$	2MASS $J-H$ color	0.59 ± 0.27	0.48 ± 0.25
32.....	$H-K_s$	2MASS $H-K_s$ color	0.32 ± 0.37	0.19 ± 0.31

表 4-2 样本参数的统计特性

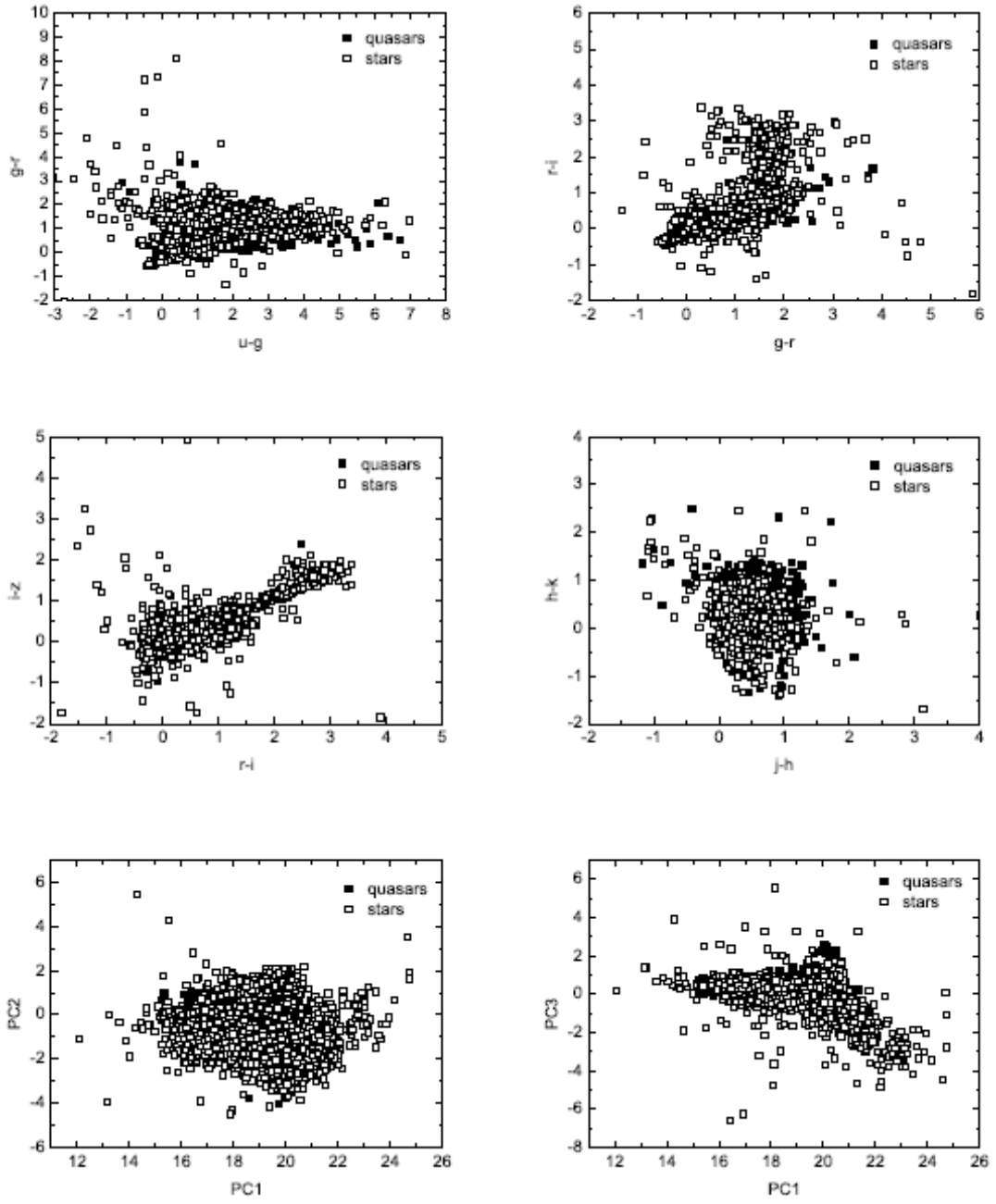


图 4-2　类星体和恒星抽样样本散点图，前 4 个是不同颜色组合的双色图，后 2 个是前三个主分量空间的分布图

4.3.3 算法性能评估量

除了总的分类准确率，我们还计算了 precision、recall、true positive rate（Acc^+）、true negative rate（Acc^-）、F-measure（FM）、G-mean（GM）以及 Weighted

Accuracy（WA）来评估分类算法的性能。这些评估量都是由表 4-3中的含混矩阵方程计算得到的。假设有两类天体分类，一类标记为正类，一类标记为负类，其中 TP 是分类正确的正类的简称，FN 是分类错误的负类的简称（即被误分为正类），而 FP 是分类错误的正类的简称，TN 是分类正确的负类的简称。在我们的分类过程中，类星体被标记为正类，恒星被标记为负类。表 4-3中的含混矩阵，行表示真实的类型，列表示被预测为的类型。基于表 4-3中的含混矩阵，上面提到的评估量由如图 4-3和图 4-4所示的方程式给出。其中，recall 是真实的正类被分类正确的比例，而 precision 是被预测的正类中预测正确的比例。对任意的分类器，recall 和 precision 之间有一个权衡。GM 对决定平均因素很有用，而 FM 可以被解释为 recall 和 precision 的权值均值。WA 使用了一个可调的参数 β 来适应不同的应用，本实验中我们使用 β =0.5。这些评估计量被广泛应用于信息检索领域来评估性能，我们采用这些计量来比较不同参数对算法结果的影响。训练－测试方法和 10 倍的交叉确认方法被用于获得所有性能的评估计量。

表 4-3 含混矩阵

	预测为正类	预测为负类
真实的正类	TP	FN
真实的负类	FP	TN

$$\text{Accuracy} \quad (\text{Acc}) = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}} \tag{1}$$

$$\text{True Positive Rate} \quad (\text{Acc}^{+}) = \frac{\text{TP}}{\text{TP} + \text{FN}} = \text{Recall} \tag{2}$$

$$\text{True Negative Rate} \quad (\text{Acc}^{-}) = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{3}$$

图 4-3 公式（1, 2, 3）

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{4}$$

$$F - measure \quad (\text{FM}) = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{5}$$

$$\text{G} - \text{mean} \quad (\text{GM}) = (\text{Acc}^{-} \times \text{Acc}^{+})^{\frac{1}{2}} \tag{6}$$

$$\text{WeightedAccuracy} \quad (\text{WA}) = \beta \times \text{Acc}^{+} + (1 - \beta) \times \text{Acc}^{-}. \tag{7}$$

图 4-4 公式（4, 5, 6, 7）

4.3.4 结果与讨论

我们的实验使用了 Simon Levy 编写的 kd-tree 算法 JAVA 包（http://www.cs.wlu.edu/levy/software/kd）和 C 语言实现的 SVMs 算法工具 SVMLight（http://svmlight.joachims.org/）。PC 电脑的配置为 Microsoft Windows XP，Pentium(R) 4，3.2GHz CPU，1.00 GB 内存。

我们研究不同输入参数对 kd-tree 和 SVMs 分类准确率的影响，并比较了这两种方法基于不同输入参数的性能。实验结果如表 4-4至表 4-8所示。

首先，采用 kd-tree 用不同输入参数来分类类星体和恒星。每个样本都被随机的分成两部分，三分之二用于训练得到分类器，三分之一用于测试得到分类准确率，这个方法就是训练-测试方法。对于不同的输入参数，样本的数量是不同的。标记 Q（Q 代表类星体）和 S（S 代表恒星）被添加到样本中用来构建有监督的 kd-tree 分类器。然后，我们用 kd-tree 找到每个测试样本在训练样本的 n 个最近邻的类别值。对每个测试样本，将在 n 个最近邻出现概率最大的类别值作为该测试样本的类别值。因此，n 必须是一个奇数来避免五-五分的情况。理论上，n 值越高越能减少训练数据中噪音的影响。而实际应用中考虑运算效率，n 的值常选择十量级而不是选择百或千量级。对 n 随机取值来实验，发现 n 取 11 时，分类准确率较高，因此在下面的试验中选 n=11。五个波段的星等（u、g、r、i、

z）先被作为 kd-tree 的输入参数，然后四个颜色（u-g、g-r、r-i、i-z）和 r 星等被作为输入参数。当选择更多参数时，可能包含的信息越多，越有助于分类，于是增加 2MASS 星表的 J、H 和 K_s 星等（或 J-H、H-K_s）作为输入来创建分类器。我们比较了基于 PSF 星等、模型星等和红化校正的模型星等的不同输入参数，比较的结果如表 4-4所示。

从表 4-4可以发现，对所有的输入参数组合，kd-tree 算法的准确率都很高，高于 94.47%，FM、GM、WA 的值也都很高，且运行时间都低于 5 分钟。一般而言，基于模型星等的输入参数组合的准确率比基于其他类型星等的要高，而基于红化校正模型星等的准确率又比基于 PSF 星等的要高。对于这三种星等，基于四个颜色和r星等作为输入参数的结果比基于五个颜色的要好。当参数更多时，准确率不一定会提高，比如当输入参数加上 2MASS 星表的参数时，准确率反而下降了。只有采用适合的参数，准确率才会最好。当输入参数较少时，kd-tree 的性能较好且建分类器的时间越少。如表 4-4中所示，当采用模型星等的四个颜色和 r 星等作为输入参数时，得到的准确率最高为 97.26%，且同时得到最高的 FM、GM 和 WA 的值分别为 96.69%、97.14%和 97.14%，运行时间也较短，不超过 1 分钟。

Input patterns	Acc (%)	Acc^+ (%)	Acc^- (%)	Precision (%)	FM (%)	GM (%)	WA (%)	Time (s)
u^p, g^p, r^p, i^p, z^p	96.32	95.34	97.02	95.76	95.55	96.17	96.18	27
$u^p - g^p, g^p - r^p, r^p - i^p, i^p - z^p, r^p$	97.02	96.24	97.58	96.56	96.40	96.90	96.91	48
$u^p, g^p, r^p, i^p, z^p, J - H, H - K_s$	95.82	94.90	96.48	95.01	94.96	95.69	95.69	83
$u^p - g^p, g^p - r^p, r^p - i^p, i^p - z^p, r^p, J - H, H - K_s$	96.62	95.89	97.14	95.96	95.92	96.51	96.52	166
$u^p, g^p, r^p, i^p, z^p, J, H, K_s$	94.81	93.91	95.45	93.58	93.75	94.67	94.68	90
$u^p - g^p, g^p - r^p, r^p - i^p, i^p - z^p, r^p, J, H, K_s$	95.76	95.08	96.24	94.70	94.89	95.66	95.66	166
u, g, r, i, z	96.46	95.42	97.20	96.02	95.72	96.31	96.31	26
u-g,g-r,r-i,i-z,r	**97.26**	96.41	97.87	96.97	**96.69**	**97.14**	**97.14**	54
$u, g, r, i, z, J - H, H - K_s$	95.85	94.90	96.53	95.09	94.99	95.71	95.72	91
$u - g, g - r, r - i, i - z, r, J - H, H - K_s$	96.76	96.02	97.28	96.15	96.08	96.65	96.65	148
u, g, r, i, z, J, H, K_s	94.87	93.88	95.57	93.75	93.81	94.72	94.73	91
$u - g, g - r, r - i, i - z, r, J, H, K_s$	95.85	95.09	96.39	94.90	95.00	95.74	95.74	167
u', g', r', i', z'	96.41	95.42	97.10	95.88	95.65	96.26	96.26	25
$u' - g', g' - r', r' - i', i' - z', r'$	97.19	96.37	97.76	96.82	96.60	97.07	97.07	44
$u', g', r', i', z', J - H, H - K_s$	95.8	95.00	96.47	95.01	95.00	95.73	95.74	81
$u' - g', g' - r', r' - i', i' - z', r', J - H, H - K_s$	96.68	95.97	97.18	96.00	95.99	96.57	96.57	151
$u', g', r', i', z', J, H, K_s$	94.73	93.78	95.41	93.52	93.65	94.59	94.59	89
$u' - g', g' - r', r' - i', i' - z', r', J, H, K_s$	95.76	95.04	96.27	94.74	94.89	95.65	95.65	178

表 4-4 使用 kd-tree 算法 n=11 时不同输入参数准确率的对比

从上面的结果可以总结出以模型星等的四个颜色和 r 星等作为输入时，kd-tree 算法在当 n=11 时获得最佳效果。接着我们来考察采用这组输入参数，不同 n 值对 kd-tree 性能的影响。通过比较准确率、FM、GM、WA 和运行时间，我们来评估不同 n 值对构建分类器的效率和效果，从而得到最佳的 n 值。我们采用从 3 到 29 的奇数作为 n 值，表 4-5显示最好的分类准确率是当 n=11 时的 97.263%，其次好的结果分别是当 n=7 和 9 时的 97.262%和 97.252%。当 n 值增加时运行时间会增加。最高的 FM、GM 和 WA 值是当 n=7 时，分别为 96.690%、97.149%和 97.151%。我们还看出当 n=7 时分类正确的类星体比例比 n=9 或 11 时要高，这表明当 n=7 时分类器对类星体有更高的预测准确率，同时还保持合理的恒星预测准确率。

n	Acc (%)	Acc^+ (%)	Acc^- (%)	Precision (%)	FM (%)	GM (%)	WA (%)	Time (s)
3	97.167	96.479	97.654	96.677	96.578	97.065	97.066	29
5	97.239	96.554	97.723	96.775	96.665	97.137	97.139	35
7	**97.262**	96.503	97.799	96.877	**96.690**	**97.149**	**97.151**	41
9	97.252	96.452	97.818	96.902	96.677	97.133	97.135	46
11	**97.263**	96.413	97.866	96.966	96.689	97.137	97.140	50
13	97.228	96.412	97.804	96.882	96.646	97.106	97.108	54
15	97.187	96.342	97.785	96.853	96.597	97.061	97.064	57
17	97.130	96.227	97.768	96.826	96.526	96.994	96.998	61
19	97.102	96.153	97.774	96.831	96.491	96.960	96.964	64
21	97.081	96.149	97.740	96.785	96.466	96.941	96.945	67
23	97.064	96.113	97.737	96.780	96.445	96.922	96.925	70
25	97.043	96.082	97.723	96.760	96.420	96.899	96.903	73
27	97.004	96.023	97.698	96.724	96.372	96.857	96.861	76
29	96.968	95.956	97.684	96.702	96.328	96.816	96.820	78

表 4-5 使用 kd-tree 算法当输入参数为 u-g、g-r、r-i、i-z、r 时不同 n 值的结果的比较

因为对 kd-tree 算法最好的输入参数是模型星等的四个颜色和 r 星等，我们也用这个输入参数来建造 SVMs 分类器。选择径向基核函数（Radial Basis Function，简称 RBF）作为 SVMs 的核函数，此时 RBF SVMs 有两个可调参数：

惩罚因子 c 和核参数 ɣ。我们尽力寻找最优化参数组合（c，ɣ），实验结果列在表 4-6中。当使用 ɣ =5 且 c =1 或 5 时获得最高的准确率（97.50%），且运行时间分别为 21 和 28 分钟。但是，当 ɣ=8 和 c=1 时获得最高的 FM 值为 97.78%，而当 ɣ=5 和 c=5 时获得最高的 GM 和 WA 分别为 97.41%和 97.41%。我们还发现分类正确的类星体比例在 ɣ=5 和 c=5 时比在 ɣ=5 和 c=1 时要高。表 4-6所示当 c 值越小时运行时间也越少，当 ɣ=5 和 c=0.1 时建 SVMs 分类器的时间最少，而当 ɣ=0.01 和 c=1000 时消耗时间高达 1 天 14 小时。

Algorithm (RBF kernel)	Soft margin c	Acc (%)	Acc^+ (%)	Acc^- (%)	Precision (%)	FM (%)	GM (%)	WA (%)	Time (m)
$\gamma = 5$	$c = 0.1$	96.85	95.14	98.07	97.22	97.64	96.59	96.60	17
$\gamma = 0.01$	$c = 1$	92.09	88.69	94.50	91.95	93.21	91.55	91.60	32
γ=5	**c=1**	**97.50**	96.67	98.09	97.27	97.68	97.38	97.38	21
$\gamma = 8$	$c = 1$	97.48	96.50	98.17	97.39	**97.78**	97.33	97.34	30
$\gamma = 0.01$	$c = 10$	92.90	89.37	95.40	93.22	94.30	92.34	92.39	47
$\gamma = 5$	$c = 10$	97.41	96.86	97.79	96.88	97.33	97.32	97.33	69
γ=5	**c=5**	**97.50**	96.88	97.93	97.07	97.50	**97.41**	**97.41**	49
$\gamma = 8$	$c = 5$	97.38	96.61	97.92	97.04	97.48	97.26	97.26	54
$\gamma = 0.01$	$c = 1000$	94.55	94.26	96.02	92.47	93.60	94.50	94.50	2280

表 4-6 使用 RBF SVMs 算法当输入参数为 u-g、g-r、r-i、i-z、r 时不同可调参数的结果的比较

考虑最好的 GM 和 WA 的值，我们认为最佳参数组合为 ɣ=5 和 c=5。接着，我们设定参数 ɣ=5 和 c=5，比较不同输入参数的性能，结果如表 4-7所示。很明显基于准确率、FM、GM 和 WA，最好的输入参数是(u-g、g-r、r-i、i-z、r)。使用模型星等的四个颜色和 r 星等作为输入参数训练 SVMs 分类器的性能比使用五个星等（u、g、r、i、z）好。和 kd-tree 的结果相似，基于模型星等的结果比基于红化校正的模型星等的好，而基于红化校正的模型星等的又比基于 PSF 星等的好。当加入更多红外波段的参数，性能并没有提高，反而降低了，这表明 J-H 和 $H-K_s$ 参数对分类效果几乎没有贡献。

Input patterns	Acc (%)	Acc^+ (%)	Acc^- (%)	Precision (%)	FM (%)	GM (%)	WA (%)	Time (m)
u^p, g^p, r^p, i^p, z^p	96.97	96.47	97.40	96.97	97.19	96.94	96.94	58
$u^p - g^p, g^p - r^p, r^p - i^p, i^p - z^p, r^p$	97.39	96.95	97.77	97.39	97.58	97.36	97.36	42
u, g, r, i, z	97.15	96.64	97.59	97.16	97.37	97.11	97.11	61
u-g,g-r,r-i,i-z,r	**97.50**	96.97	97.93	97.49	**97.71**	**97.45**	**97.45**	44
$u-g, g-r, r-i, i-z, r, J-H, H-K_s$	97.17	96.16	97.93	97.18	97.55	97.04	97.04	62
u', g', r', i', r'	97.15	96.72	97.53	97.16	97.34	97.12	97.12	60
$u'-g', g'-r', r'-i', i'-z', r'$	97.47	97.02	97.85	97.47	97.66	97.44	97.44	43

表 4-7 使用 RBF SVMs 算法当参数 c=5、γ=5 时不同输入参数的结果的比较

最后，我们使用 10 倍的交叉确认方法来评估 kd-tree 和 SVMs 算法的性能和建造分类器的速度，并采用模型星等的四个颜色和 r 星等作为输入参数。交叉确认是一种通用且有效的技术用于机器学习常面临的很多问题，比如准确率评估、参数选择或参数调优。交叉确认方法被广泛用于机器学习的方法，比如 kd-tree 和 SVMs。K 倍交叉确认是一种重要的交叉确认方法，数据被随机分成 k 个子样本，每一次用一个子样本作为测试数据，另外 k-1 个子样本放到一起形成训练数据，循环 k 次，然后计算平均正确率及其误差。我们用交叉确认方法来比较两个算法的效率和效果。如表 4-8所示，关于准确率的最好结果是当 RBF 核函数的 SVMs 分类器的参数 γ=5 与 c=5 时，分类准确率为 97.65%且方差为 0.22%；当 γ=2 和 c=10 时，准确率为（97.65±0.22）%；当 γ=5 和 c=1 时，准确率为（97.64±0.20）%，其中最后情况的方差是三种情况中最小的。最高的 FM 值是当 γ=10 和 c=1 时的 98.01%；在 γ=5 和 c=5 时，GM 和 WA 取得最高值，分别为 97.55%和 97.56%。当 n=9 时，kd-tree 得到的结果最好，最高的 FM、GM 和 WA 的值分别为 97.45%、97.66%和 97.32%，且这些值比当 n=11 时只稍高一点。基于表 4-8，我们很难评出 kd-tree 和 SVMs 的优劣。不过，这两种算法的准确率都高于 97.0%，因此是分类类星体和恒星的有效分类器。

$n/(\gamma,c)$	Acc (%)	Acc^+ (%)	Acc^- (%)	Precision (%)	FM (%)	GM (%)	WA (%)
$n=13$	97.40 ± 0.23	96.47 ± 0.31	98.06 ± 0.34	97.24 ± 0.48	97.65 ± 0.41	97.26 ± 0.23	97.26 ± 0.23
$n=11$	97.42 ± 0.25	96.53 ± 0.31	98.06 ± 0.35	97.24 ± 0.48	97.65 ± 0.41	97.29 ± 0.24	97.29 ± 0.24
n=9	**97.45±0.23**	96.58 ± 0.28	98.07 ± 0.34	97.26 ± 0.47	**97.66±0.41**	**97.32±0.22**	**97.32±0.22**
$n=7$	97.44 ± 0.23	96.60 ± 0.30	98.03 ± 0.34	97.21 ± 0.47	97.62 ± 0.41	97.31 ± 0.22	**97.32±0.22**
c=1, γ=5	**97.64±0.20**	96.76 ± 0.26	98.26 ± 0.34	97.52 ± 0.47	97.89 ± 0.40	97.50 ± 0.19	97.51 ± 0.19
$c=1,\gamma=8$	97.62 ± 0.21	96.64 ± 0.22	98.31 ± 0.33	97.59 ± 0.46	97.95 ± 0.40	97.47 ± 0.19	97.47 ± 0.19
$c=1,\gamma=10$	97.59 ± 0.22	96.50 ± 0.26	98.37 ± 0.34	97.66 ± 0.48	**98.01±0.41**	97.43 ± 0.21	97.43 ± 0.21
$c=1,\gamma=2$	97.46 ± 0.19	96.63 ± 0.26	98.04 ± 0.32	97.22 ± 0.43	97.63 ± 0.38	97.33 ± 0.17	97.34 ± 0.17
c=5, γ=5	**97.65±0.22**	96.97 ± 0.24	98.14 ± 0.35	97.37 ± 0.35	97.75 ± 0.42	**97.55±0.20**	**97.56±0.20**
$c=5,\gamma=8$	97.61 ± 0.26	97.40 ± 0.28	98.17 ± 0.39	97.40 ± 0.54	97.78 ± 0.46	97.49 ± 0.24	97.50 ± 0.24
$c=5,\gamma=10$	97.54 ± 0.26	96.66 ± 0.31	98.17 ± 0.37	97.39 ± 0.52	97.78 ± 0.45	97.41 ± 0.25	97.41 ± 0.25
$c=10,\gamma=5$	97.61 ± 0.26	96.96 ± 0.27	98.07 ± 0.41	97.27 ± 0.56	97.67 ± 0.48	97.51 ± 0.24	97.51 ± 0.24
$c=10,\gamma=8$	97.50 ± 0.26	96.74 ± 0.30	98.03 ± 0.40	97.20 ± 0.55	97.61 ± 0.48	97.38 ± 0.25	97.39 ± 0.25
$c=10,\gamma=10$	97.40 ± 0.25	96.53 ± 0.32	98.02 ± 0.36	97.19 ± 0.50	97.60 ± 0.43	97.27 ± 0.24	97.28 ± 0.24
c=10,γ=2	**97.65±0.22**	97.00 ± 0.28	98.10 ± 0.35	97.31 ± 0.49	97.70 ± 0.42	**97.55±0.20**	97.55 ± 0.20

表 4-8 使用 10 倍交叉确认方法当输入参数为 u-g、g-r、r-i、i-z、r 时不同算法不同可调参数的结果的比较

从上面表 4-4到表 4-8中得出 kd-tree 和 SVMs 算法在分类类星体和恒星的准确率方面是比较高的。当考虑运行时间时，kd-tree 要比 SVMs 快很多，图中 kd-tree 是以秒为计量单位，而 SVMs 是以分钟为计量单位。考虑准确率和速度，kd-tree 比 SVMs 要好，因为建造 SVMs 分类器的速度非常慢。而且，使用 10 倍的交叉确认方法得到的结果比使用训练－测试方法好，因为交叉确认方法可以产生一个有效的无偏差的误差估计，但是它花费的计算时间较长，是训练－测试方法的 10 倍以上。从 kd-tree 和 SVMs 的分类效率而言，由 kd-tree 和 SVMs 训练得到的分类器可以较准确地用于预测未知天体的类型，并应用于为 LAMOST 或其他巡天星表预选类星体候选体的工作中。

使用 kd-tree 和 SVMs 算法错分的源是因为其自身的内在属性特征，实验结果证明大多数用 kd-tree 错分的源和用 SVMs 错分的源是重叠的。为了可视化分类结果，我们以使用 kd-tree 为例，图 4-5为类星体和错分类星体的红移分布情况。图 4-5中所示，类星体样本的红移高峰值是在红移 1 至 2 的区域，而错分类星体的红移高峰值是在红移 2.5 至 4 的区域。错分类星体红移分布的最高峰值在红移值大约 2.8 的地方，这正是使用 SDSS 测光系统来区分 M 型恒星和类星体开

始有问题的红移区域。错分说明，无论使用什么分类方法，因为天体本身的属性导致了同样的偏差。图 4-6所示错分类星体 r 波段模型星等的峰值在星等较暗的地方，这是因为高红移的光谱观测样本的星等限制较暗这个事实。接着，我们研究了分类结果中星等的影响，如图 4-6所示，错分类星体或恒星的峰值比类星体和恒星的星等更暗。这就是说，更暗的天体更容易错分，也许因为暗源的样本数量小和信噪比低。为了进一步研究为什么错分的源易于错分，我们将错分的类星体和恒星在 SIMBAD 和 NED 中查阅。发现错分的恒星大部分是 CV 恒星、白矮星、RR 型星、碳星，还有一些是紫外源、X 射线源、射电源、蓝星，还有一些天体是星系和不规则漩涡和少量类星体。893 个错分的类星体大部分是暗星等的源，一些是 AGN、Seyfert 1、Seyfert 2 和射电源，171 个是为证认的类星体，4 个白矮星，1 个 CV 星，1 个 AM 星和 29 星系。

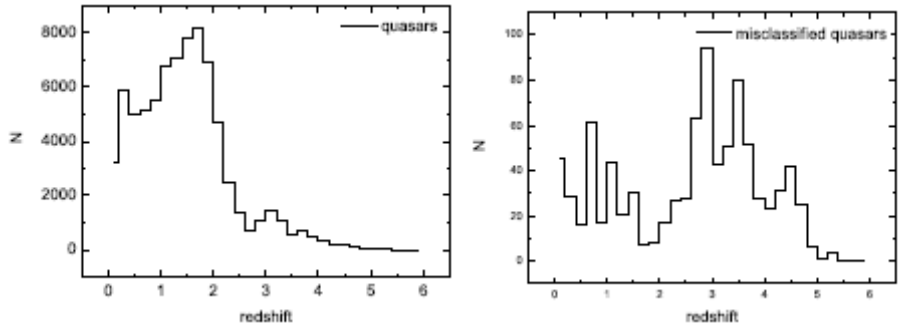


图 4-5 类星体和错分类星体的红移分布

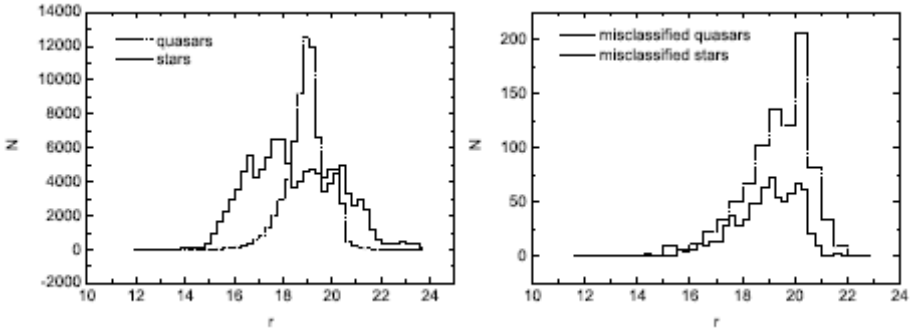


图 4-6 类星体和恒星及错分的类星体和恒星的模型 r 星等分布

4.4 用随机森林算法分类类星体、恒星和星系

本节采用不同的多波段数据，如：可见光和射电波段数据、可见光和 X 射线波段，将随机森林应用于类星体与恒星的分类，类星体、恒星和星系的分类。并比较了基于特征选择、特征加权和无特征处理的样本的分类效果。

4.4.1 巡天

SDSS（the Sloan Digital Sky Survey）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[81] [82] 。该巡天已在 **4.3.1** 节介绍，这里不再累述。

USNO-B1.0 (the United States Naval Observatory Astrometry)星表是覆盖全天的巡天表[86] ，多种巡天应用 7,435 片施密特底片经历了长达 50 年的时间拍摄得到的，其来自 3,648,832,040 次的独立观测，源的数目达到 1,045,913,669。该星表包含了天体的位置、自行、各个波段的星等、恒星/星系的估计。覆盖极限星等达到 V=21，历元 2000 年的位置精度可以达到 0.2 角妙，5 个颜色的精度达到 0.3 等，区别恒星和星系的准确率为 85%。

FIRST 巡天 (the Faint Images of the Radio Sky at Twenty-cm Survey，简称 FIRST)是一个正在运行的射电快照巡天[84] ，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。

ROSAT 巡天是 X 射线波段的巡天项目[85] 。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（Ldetect）。

4.4.2 样本与参数

基于 XMaS_VO 交叉证认多波段巡天数据，我们得到的多波段数据包括两个样本：样本 1 和样本 2。如表 4-9所示，样本 1 是包括 SDSS DR5、FIRST 和 USNO-B1.0 星表参数的多波段数据，包括 6,479 个类星体、785 个恒星和 27,984 个星系。样本 2 是包括 SDSS DR5、ROSAT 和 USNO-B1.0 星表参数的多波段数据，包括 6,360 个类星体、2,357 个恒星和 5,333 个星系。

表 4-9 样本描述

样本	星表	类星体数量	恒星数量	星系数量
样本 1	SDSS, FIRST, USNO	6,479	785	27,984
样本 2	SDSS, ROSAT, USNO	6,360	2,357	5,333

我们选择 SDSS DR5 中 5 个经过星系消光红化校正的模型星等（u、g、r、i、z）和 4 个颜色（u-g、g-r、r-i、i-z）作为输入参数，并从 FIRST 巡天中提取参数 $\log(f_{peak})$、$\log(f_{int})$、f_{maj}、f_{min}、f_{pa}，从 ROSAT 星表中提取参数 log(CR)、HR1、HR2、ext、extl，以及从 USNO-B1.0 星表中提取 B 和 R 星等作为输入参数。对样本 1 选择的输入参数包括：u-g、g-r、r-i、i-z、r、$\log(f_{peak})$、$\log(f_{int})$、f_{maj}、f_{min}、

f_{pa}、z+2.5log(f_{peak})、 z+2.5log(f_{int})、B-R；对样本 2 选择的输入参数包括：u-g、g-r、r-i、i-z、r、log(CR)、HR1、HR2、ext、extl、u+2.5log(CR)、B-R。其中参数 z+2.5log(f_{peak})和 z+2.5log(f_{int})可以被看作是射电到光学的流量比，而参数 u+2.5log(CR)可以被看作是 X 射线到光学的流量比。

表 4-10和表 4-11中总结了样本 1 和样本 2 的统计特性。从表 4-10和表 4-11中我们可以很清楚的看出星系的特性和恒星的很相似，这可能是因为星系是由恒星组成的这一事实，但类星体的特性与恒星及星系的差别则很大。基于不同类型天体参数属性的差异，显示出我们选择的参数是合理的、可行的，均有助于分类。可以将这些参数作为分类器的输入应用于分来类星体、恒星和星系，尤其是应用于区分类星体和恒星。

表 4-10 样本 1 参数统计特性

参数	星表	类星体	恒星	星系
u-g	SDSS	0.56±0.76	1.89±0.79	1.85±0.84
g-r	SDSS	0.30±0.36	0.97±0.46	1.17±0.44
r-i	SDSS	0.19±0.22	0.58±0.45	0.53±0.19
i-z	SDSS	0.11±0.18	0.30±0.26	0.34±0.13
r	SDSS	18.70±1.11	18.38±1.45	17.32±1.56
log(f_{peak})	FIRST	0.90±0.72	0.49±0.47	0.40±0.38
log(f_{int})	FIRST	0.92±0.75	0.58±0.51	0.48±0.43
f_{maj}	FIRST	2.11±2.19	3.80±3.64	3.84±3.36
f_{min}	FIRST	0.64±1.00	1.37±2.13	1.10±1.70
f_{pa}	FIRST	98.01±56.41	88.77±54.47	89.44±54.19
z+2.5log(f_{peak})	SDSS, FIRST	20.65±2.16	18.71±1.65	17.46±1.67
z+2.5log(f_{int})	SDSS, FIRST	20.70±2.23	18.94±1.71	17.67±1.72
B-R	USNO-B1.0	-0.48±1.33	-0.64±3.27	-0.82±3.65

表 4-11 样本 2 参数统计特性

参数	星表	类星体	恒星	星系
u-g	SDSS	0.32±0.41	1.68±1.07	1.68±0.90
g-r	SDSS	0.22±0.29	0.90±0.79	1.02±0.47
r-i	SDSS	0.18±0.21	0.51±0.71	0.45±0.30
i-z	SDSS	0.14±0.21	0.25±0.62	0.31±0.29
r	SDSS	18.18±1.08	18.58±1.76	17.66±1.65
log(CR)	ROSAT	-1.46±0.33	-1.48±0.34	-1.44±0.39
HR1	ROSAT	-0.09±0.53	0.05±0.61	0.22±0.58
HR2	ROSAT	0.11±0.58	0.10±0.56	0.17±0.52
ext	ROSAT	3.77±8.24	4.70±11.44	8.96±21.11
extl	ROSAT	0.46±2.93	0.83±5.83	9.14±126.25
z+2.5log(CR)	SDSS, ROSAT	15.06±1.05	17.47±2.70	16.76±2.20
B-R	USNO-B1.0	-0.43±1.13	-0.34±3.24	-0.96±2.67

我们首先用随机森林分类器分别基于样本 1 和样本 2 来对类星体和恒星进行分类。随机森林可以计算每个属性的参数权值，因此该方法可以用于实现参数加权和参数选择，我们分别应用于样本 1 和样本 2。表 4-12和表 4-13分别显示了两个样本中每个属性的权值。样本 1 中权值重要的属性顺序如下：u-g、r-i、i-z、g-r、r、f_{maj}、f_{min}、z+2.5log(f_{peak})、z+2.5log(f_{int})、B-R、log(f_{peak})、log(f_{int})、f_{pa}。样本 2 中权值重要的属性顺序如下：u-g、r-i、i-z、g-r、u+2.5log(CR)、r、log(CR)、B-R、HR1、HR2、extl、ext。在随后的实验中，参数权值就是采用表 4-12和表 4-13中所示的值，而参数选择是选择了参数权值中权值较大的 7 个参数，即样本 1 参数选择的属性为：u-g、r-i、i-z、g-r、r、f_{maj}、f_{min}，样本 2 参数选择的属性为：u-g、r-i、i-z、g-r、u+2.5log(CR)、r、log(CR)。

表 4-12 样本 1 的参数权值

参数	参数权值
u-g	142.353
r-i	96.731
i-z	67.797
g-r	60.550
r	44.559
f_{maj}	36.147
f_{min}	30.680
$z+2.5\log(f_{peak})$	20.998
$z+2.5\log(f_{int})$	19.061
B-R	16.725
$\log(f_{peak})$	12.867
$\log(f_{int})$	10.668
f_{pa}	1.535

表 4-13 样本 2 的参数权值

参数	参数权值
u-g	196.366
r-i	81.819
i-z	59.084
g-r	48.853
u+2.5log(CR)	43.924
r	35.813
log(CR)	29.164
B-R	18.859
HR1	11.293
HR2	8.649
extl	3.871
ext	1.241

4.4.3 结果与讨论

我们用 10 倍的交叉确认方法对所有参数、带权值的参数和参数选择的参数分别进行了类星体和恒星的分类，实验结果如表 4-14和表 4-15所示。基于样本 1 的三种参数情况，类星体的分类准确率都高于 99.00%，而恒星的分类准确率都高于 84.00%，总的准确率都高于 97.60%。基于样本 2 的三种参数情况，类星体的分类准确率都高于 98.00%，恒星的分类准确率都高于 96.00%，总的准确率都高于 98.00%。表 4-14和表 4-15显示了基于样本 2 的结果比基于样本 1 的好，这就是说，使用基于光学和 X 射线波段信息比使用基于光学和射电波段信息更容易分类类星体和恒星。表 4-14和表 4-15也显示了使用带权值参数和参数选择的参数比使用所有参数的结果稍好一点，这是因为随机森林在建模过程中能挑出最有用的参数。因此，如果有很多输入参数，我们一般不需要在建模前做参数选择工作。

表 4-14 用样本 1 分类类星体和恒星的结果

样本	所有参数		带权值参数		参数选择的参数	
分类\已知	类星体	恒星	类星体	恒星	类星体	恒星
类星体	6430	119	6431	118	6423	110
恒星	49	666	48	667	56	675
准确率（%）	99.24±0.20	84.84±3.57	99.26±0.22	84.95±3.58	99.14±0.37	85.99±3.70
总准确率（%）	97.69±0.44		97.71±0.43		97.72±0.40	

表 4-15 用样本 2 分类类星体和恒星的结果

样本	所有参数		带权值参数		参数选择的参数	
分类\已知	类星体	恒星	类星体	恒星	类星体	恒星
类星体	6281	87	6281	87	6297	86
恒星	79	2270	79	2270	63	2271
准确率（%）	98.76±0.61	96.31±0.99	98.76±0.61	96.31±0.99	99.01±0.60	96.35±1.07
总准确率（%）	98.10±0.51		98.10±0.51		98.29±0.52	

随后，我们用随机森林分类器分别基于样本 1 和样本 2 对类星体、恒星和星系进行分类。基于表 4-14和表 4-15的实验结果，三种参数情况的准确率几乎一样，因此我们只考虑所有参数的情况。基于所有参数分类类星体、恒星和星系的结果如表 4-16和表 4-17所示。所有类型的天体中，恒星的准确率最低，这可能是因为恒星的样本数最少。基于样本 1 和样本 2 的总的准确率分别为 96.17%和 91.08%。很明显，基于样本 1 的结果比基于样本 2 的结果好，这说明使用基于光学和射电波段信息比使用基于光学和 X 射线波段信息更容易来分类类星体、恒星和星系。

表 4-16 用样本 1 分类类星体、恒星和星系的结果

分类\已知	类星体	恒星	星系
类星体	5783	70	364
恒星	15	516	33
星系	681	199	27587
准确率（%）	89.26±1.47	65.732±4.79	98.58±1.65
总准确率（%）	97.17±0. 30		

表 4-17 用样本 2 分类类星体、恒星和星系的结果

分类\已知	类星体	恒星	星系
类星体	6005	67	409
恒星	40	2055	188
星系	315	235	4736
准确率（%）	94.42±1.12	87.19±2.54	88.81±1.57
总准确率（%）	91.08±0. 92		

4.5 用 kd-tree 算法评估星系的测光红移

红移的测量对于研究星系的形成和演化以及宇宙大尺度结构具有重要的意义，目前已有多种数据挖掘方法（如：核回归、支持矢量机、随机森林等）应用于测光红移的估测。这里我们将已知光谱红移的 SDSS 星系测光数据样本分成训练集和测试集。用训练集训练 kd-tree，获得回归器，由测试集来评估该回归器用于红移测量的效果。

4.5.1 样本

我们选取了经过光谱观测的 SDSS DR5 的星系测光数据。首先对数据进行去噪和去缺值处理。为了评估测光红移，样本的选择是按 2004 年 Vanzella 给出的标准[87] ，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）。符合该标准的有 375,929 个星系，我们把这些星系数据随机地分成训练集（251,872 个）和测试集（124,057 个）。在本实验中，我们直接使用了红化校正星等的四个颜色（u-g、g-r、r-i、i-z）作为输入参数，然后再把红化校正星等的四个颜色和 r 星等一起作为 kd-tree 的输入参数。

4.5.2 结果与讨论

表 4-18给出了用于评估测光红移的每个参数集的误差估计。当选择四个颜色作为输入时，测试数据的评估误差为 0.0212；当选择四个颜色和 r 星等一起作为输入时，评估误差增至 0.0232。这表明，仅使用四种颜色作为输入比使用四种颜色以及 r 星等的情况，kd-tree 算法评估测光红移的性能更优越。李丽丽等人用神经网络（Artificial Neural Networks，简称 ANNs）算法对 SDSS 星系数据的红化校正星等进行了测光红移研究[88] ，当选择四个颜色及 r 星等作为输入参数时，测试数据的评估误差为 0.021079。由此可见，kd-tree 和 ANNs 在预测红移时效果相当。可是在进行算法比较时，由于各种方法所用样本和参数的差异，致使我们只能给出比较粗糙的对比，见表 4-18和表 4-19（该表引自王丹博士论文[89]第 71 页）。对比发现 kd-tree 明显优于模板匹配方法（如 CWW、Bruzual-Charlot）、CMR、Interpollated、多项式回归方法（如 Polynomial、MPR）和 SVMs，与 ANNs

获得的结果相当，精度要低于核回归的结果。但是核回归具有在实施时存在损失点而且预测速度较慢等特点，因此考虑效率和效果时，kd-tree 方法是不错的选择。将来的工作可以考虑将核回归与 kd-tree 相结合，综合二者的优点，将大大提高红移预测的效率和效果。

表 4-18 用 kd-tree 算法评估星系测光红移

输入参数	**评估误差**
红化校正星等 u-g, g-r, r-i, i-z	0.0212
红化校正星等 u-g, g-r, r-i, i-z, r	0.0232

表 4-19 不同测光红移方法的对比

方法	剩余标准偏差$\sigma_{\rm rms}$	数据样本	输入参数
CWW[1]	0.0666	SDSS-EDR	$ugriz$
Bruzual-Charlot[1]	0.0552	SDSS-EDR	$ugriz$
Bayesian[2]	0.0476	HDF-N	$UBVIJHK$(PCA)
CMR[3]	0.0320	SDSS DR4	$ugriz$
Interpolated[1]	0.0451	SDSS-EDR	$ugriz$
Polynomial[1]	0.0318	SDSS-EDR	$ugriz$
MPR[3]	0.0256	SDSS DR5	color
Kd-tree[1]	0.0254	SDSS-EDR	$ugriz$
ClassX[4]	0.0340	SDSS-DR2	$ugriz$
SVMs[5]	0.027	SDSS-DR2	$ugriz$
	0.0230	SDSS-DR2	$ugriz$+petroR50+petroR90
SVMs[3]	0.0273	SDSS DR5,2MASS	color
ANNs[6]	0.0229	SDSS-DR1	$ugriz$
Polynomial[7]	0.025	SDSS-DR1,GALEX	$ugriz+nuv$
KR[3]	0.0215	SDSS-DR5	$ugriz$
	0.0206	SDSS-DR5	color+r
	0.0189	SDSS-DR5	color+eClass
	0.0193	SDSS-DR5,2MASS	color

4.6 小结

本章从天文中的分类和回归研究谈起，不仅阐述了这两种数据挖掘任务的原理和方法，而且描述了一些与此相关的具体的天文应用。接着介绍了 kd-tree、SVMs 和随机森林算法的原理及其在天文学中的应用实例，最后具体描述了这三种算法的具体应用。针对将类星体从恒星或恒星和星系的样本中分离出来，采用了监督的 kd-tree、支持矢量机（SVMs）和随机森林（RF）算法；针对星系的测光红移估测，应用了 kd-tree 方法。

比较了两个自动分类算法：kd-tree 和支持向量机（SVMs），用于在 SDSS 和 2MASS 星表数据库中将类星体从恒星中分离出来。用 SDSS 和 2MASS 交叉证认样本中的已知光谱的数据来训练这两种方法。一方面，固定模型的可调参数，比较了不同参数组合作为输入时分类器的分类效果；另一方面，固定输入参数，考察这两个模型的可调参数的最优值。为了评价这两个算法的分类效率及模型参数的调优，我们计算了一些评估量 precision、recall、true positive rate、true negative rate、F-measure、G-mean 以及 Weighted Accuracy，以此来选取最优输入参数组合及模型的最优参数。研究结果表明 kd-tree 和 SVMs 用于进行点源分类是可靠的、有效的。对分类效率而言，SVMs 的准确率稍高，而 kd-tree 的速度明显要快。对比不同的输入参数组合，在较少的参数的情况下，kd-tree 和 SVMs 均有更好的准确率。当使用基于 SDSS 的模型星等的四个颜色（u-g、g-r、r-i、i-z）和 r 星等作为输入参数的时候，得到的准确率最高。

另外，使用随机森林（Random Forest）算法分别基于光学、射电波段数据和光学、X 射线波段数据，来研究类星体和恒星的分类，以及类星体、恒星和星系的分类。通过提取不同波段的数据，如 SDSS DR5、USNO-B1.0、FIRST 和 ROSAT，经过交叉证认得到多波段数据。随机森林算法可以实现特征选择和特征加权，于是针对我们的样本给出了每个特征的权重，从而确定出哪些参数对分类更有效。然后研究了三种输入参数（特征选择、特征加权、全部特征）情况下的分类效率，结果表明在三种情况下的分类效率相当。而且分类准确率都高于

90%。

同一算法有时即可以用于分类任务，也可以用于回归任务，我们使用的这几种算法即可以这样。而且别的工作者已经作过这方面的研究。我们利用 kd-tree 算法，基于 SDSS DR5 的星系的测光数据和光谱红移来进行测光红移估测，将红化校正星等和光谱红移作为输入参数。对比了四个颜色和四个颜色及 r 星等的两种参数组合，结果表明四个颜色的结果更好，其测试数据的预测误差为 0.0212。

综上所述，kd-tree、SVMs、随机森林用于解决天文中的分类和回归时表现出较好的性能，因此用这些算法训练后得到的分类器和回归器可用来解决虚拟天文台面临的自动分类和自动回归问题。对于分类问题，可用于各种分类任务（如星系形态的分类、恒星光谱的分类等），也可以为大型的巡天项目选取输入星表，以提高望远镜的观测效率。对回归问题，可以用于天体物理参数的测定，如星系或类星体的测光红移的测定、恒星物理参数的测量等。

第 5 章 数据挖掘应用——为 LAMOST 预选类星体候选体

大天区多目标光纤光谱望远镜（the Large Sky Area Multi-Object Fiber Spectroscopic Telescope，简称 LAMOST），是 1997 年由国家批准的大科学工程项目。在 2008 年底 LAMOST 即将完工阶段，为了高效地利用望远镜，研究和探讨 LAMOST 的输入星表具有紧迫而重要的科学意义，尤其是类星体的输入星表。利用目前已有的大型项目（如 SDSS、2MASS、NVSS、FIRST、ROSAT、USNO 等）的巡天数据，如何从中高效地选取类星体候选体，仍然是一个亟待解决的问题。上一章研究了 kd-tree、SVMs、随机森林等数据挖掘算法在天文数据挖掘中的应用，验证了这些算法的有效性和可用性。我们可以将已有光谱认证的数据的分类算法的研究成果，用于对未知天体类型的数据进行预测，从而挑出感兴趣的源，作为 LAMOST 项目的输入星表数据。

5.1 LAMOST 项目简介

LAMOST 项目于 2001 年 8 月开工建设。目前项目已经进入装配和调试阶段，并于 2007 年 6 月份 LAMOST 小系统出光（相当于 2m 口径的望远镜），整个 LAMOST 望远镜将于 2008 年底建成，开始试观测。2009 年之后就可以对类星体进行观测。LAMOST 计划观测 1 千万恒星、1 千万星系和 1 百万类星体的光谱。要从已有的海量天文数据中，挑选出满足 LAMOST 观测要求的巨量输入星表，这就需要数据挖掘技术。

LAMOST 的基本参数如下：

有效通光口径： 4 米

焦距： 20 米

视场角直径： 5 度（焦面线直径：1.75 米）

光学质量： 80%光能量集中在 2.0 角秒直径的圆内

光纤数： 4000 根

光谱覆盖范围： 370-900 纳米

观测天区： 赤纬从-10 度到+90 度的 24000 平方度

光谱分辨率： 1-0.25 纳米

光谱观测能力： 以 1 纳米的光谱分辨率，在 1.5 小时曝光时间内，极限星等达到 20.5 等

鉴于 LAMOST 望远镜的地理位置及基本参数，所选天体目标必须亮于 20.5 等，而且位置坐标须满足赤纬处于-10 度到+90 度之间。

5.2 各种预选源方法及其成功率

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，人们花费了大量的时间和精力。为了提高效率，人们采取了很多种预选源方法，如紫外超、无缝光谱、光变、零自行、多色测光及多波段交叉证认、自动化分类和聚类等方法。这些方法都各有优劣[91]

各种预选源方法及其成功率：

基于紫外超选的成功率：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 $m_b \approx 16$ 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%。

基于多色选的成功率：Warren 等人利用 SDSS(the Sloan Digital Sky Survey)巡天数据的四维颜色空间来选源[92] 。选取 106 个点源的类星体候选体，约 105 个（大部分是最亮的，占到~10%）通过光谱观测得到证认。

基于 X 射线选的成功率：EMSS(the Einstein Medium Sensitivity Survey)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%。Piccinotti 巡天发现的 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均为 I 型塞弗特星系。

基于射电选的成功率：在 FBQS 巡天中 Gregg 等人[93] 和 White 等人[94] 用射电选源的方法，成功率可以达到约 65%。

基于 γ 射线选的成功率：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。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%。

基于多种方法选的成功率：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。Brunzendorf 和 Meusinger[95] 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[96] 指出为增加样本的完备性，最好与色选择方法结合使用。

基于物理参数截断选的成功率：魏建彦等人[97] 以 $\log C \geq 0.4R+4.9$ 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%。

基于多波段选的成功率：何香涛等人[98] 采用多波段的类星体巡天（the Multiwavelength quasar survey，MWQS）方法，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[99] 在何香涛等人[98] 的工作基础上，修改了他们在 X 射线波段的选源标准，利用修改后的选择标准得到 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）。

数据挖掘预选源方法及其成功率：数据挖掘和知识发现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。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、有潜在作用的，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[100] 。具体到天文学中，数据挖掘和知识发现（Data Mining and Knowledge Discovery from Astronomical Databases）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[101] 。目前已有多种数据挖掘方法用于预选源研究，如：支持矢量机[70] 、学习矢量量化[70] 、多层神经网络[49] 、决策树（REPTree、 Random Tree、 Decision Stump、 Random Forest、J48、NBTree、 AdTree、C4.5）[50] 、决策表[102] 、统计学习（improved CHAID、 linear discriminate analysis、 Naive Bayes ）[103] 。这些工作都是基于多波段数据的样本来试验的，如交叉证认了 2MASS、ROSAT 与 USNO 星表，2MASS 与 FIRST 星表，并且将这些交叉证认后的星表与已知的类星体的星表（Veron）、星系星表（RC3）、恒星星表（SIMBAD 或 Tycho）再次交叉证认，从而获得多波段的数据样本。将样本分为两部分：一部分训练，一部分检验，发现正确率都在 90% 以上。

综上所述，目前用通常的方法选取活动星系核候选体的成功率约为 40%－80%。而且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、效率又不高。而从数据挖掘方法实验的结果来看，其不仅成功率较高，可达 90%

以上，而且适用于大样本的研究。因此我们有必要用数据挖掘的方法高效地选取预观测目标，从而提高望远镜的观测效率。

5.3 基于 SDSS 巡天数据的 LAMOST 预选源工作

SDSS (the Sloan Digital Sky Survey，简称 SDSS)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。使用设在美国的 Apache Point 天文台（APO）的 2.5 米望远镜，视场为 3 度（7 平方度），660 根光纤，光谱覆盖范围在 390-910nm，光谱分辨本领为 R~2000。该计划于 1999 年开始试观测，于 2005 年完成第一期观测，目前正在进行第二期观测。第一期目标集中在星系和宇宙学的观测，巡天范围主要是围绕北银冠 8000 平方度 5 种颜色的测光，其星系红移的巡天极限星等为 r=18.7 等，共获得 675,000 个星系、90,000 个类星体和 18,500 个恒星的光谱。第二期中 SDSS 巡天除了继续完成 I 期未观测的天区外，扩展到了恒星、银河系结构（SEGUE）和超新星巡天，SEGUE 计划在北银冠外选择 3500 平方度进行测光观测，并对 240,000 个恒星进行光谱观测，其视向速度测量精度为 10km/s，Δ[Fe/H]~0.3dex，巡天已于 2005 年开始，计划于 2008 年完成。详情参见：http://www.sdes.org/。SDSS 图像和光谱数据的信息参看表 5-1和表 5-2：

表 5-1 SDSS DR6 图像数据

<table>
<tr><td>天区覆盖</td><td>9583 平方度</td></tr>
<tr><td>图像星表</td><td>287 百万个独立源</td></tr>
<tr><td>数据量</td><td><table><tr><td>图像数据</td><td>10.0 TB</td></tr><tr><td>星表数据 (DAS, fits 格式)</td><td>2 TB</td></tr><tr><td>星表数据 (CAS, SQL 数据库)</td><td>4 TB</td></tr></table></td></tr>
<tr><td>平均波长及星等极限
(点源 95%重复探测率)</td><td><table><tr><td>u</td><td>g</td><td>r</td><td>i</td><td>z</td></tr><tr><td>3551Å</td><td>4686Å</td><td>6165Å</td><td>7481Å</td><td>8931Å</td></tr><tr><td>22.0</td><td>22.2</td><td>22.2</td><td>21.3</td><td>20.5</td></tr></table></td></tr>
<tr><td>PSF 宽度</td><td>1.4" 在 r 波段的中值</td></tr>
<tr><td>测光定标</td><td><table><tr><td>r</td><td>u-g</td><td>g-r</td><td>r-i</td><td>i-z</td></tr><tr><td>2%</td><td>3%</td><td>2%</td><td>2%</td><td>3%</td></tr></table></td></tr>
<tr><td>天测精度</td><td>< 0.1" rms 对每一坐标的绝对精度</td></tr>
</table>

表 5-2 SDSS DR6 光谱数据

天区覆盖	7425 平方度.
波长覆盖	3800-9200Å
分辨率	1800-2200
信噪比	>4 per pixel 在 g=20.2
红移精度	30 km/sec rms 对主星系样本(重复观测)
主要样本的星等极限	星系： Petrosian r <17.77 类星体：PSF i <19.1（20.2（红移>2.3））
光谱星表	1,271,680 条光谱, 其中 790,860 星系 90,108 类星体 (红移<2.3) 13,539 类星体 (红移>2.3) 218,019 恒星 69,052 M 恒星及晚型星 68,770 天光光谱 21,332 未知源

从 SDSS 中为 LAMOST 选取类星体候选体，首先源的图像表现形式应当是点源，而且应当排除掉已有光谱观测的源，更重要的是要排除掉恒星的污染。下面三个图分别描述了 SDSS 巡天中的所有点源随模型 r 星等的统计分布，没有光谱观测的点源随模型星等 r 的统计分布，以及具有光谱观测的点源随模型 r 星等的统计分布。由这三个图可以看出具有光谱观测的源占的比例相当小，因此我们可以从不具有光谱观测的源中选取 LAMOST 的预观测目标。为了提高效率，需要排除掉恒星的污染。我们可以借助监督的分类方法，具体步骤如下：选取具有光谱的恒星和类星体数据作为已知样本，对模型进行训练，得到分类器。为了检验分类方法的结果，得到分类器后，先对 SDSS 挑选出来预选类星体候选体的星表 QsoTarget 进行预测，分析预测类星体的结果，方案如图 5-4所示。然后再对

没有光谱观测的点源表预测，预测为类星体的挑选出来，再根据 LAMOST 的地理位置及观测的极限星等加以限制，这样就得到了 LAMOST 的基于 SDSS 巡天的测光数据的类星体输入星表，方案如图 5-5所示。

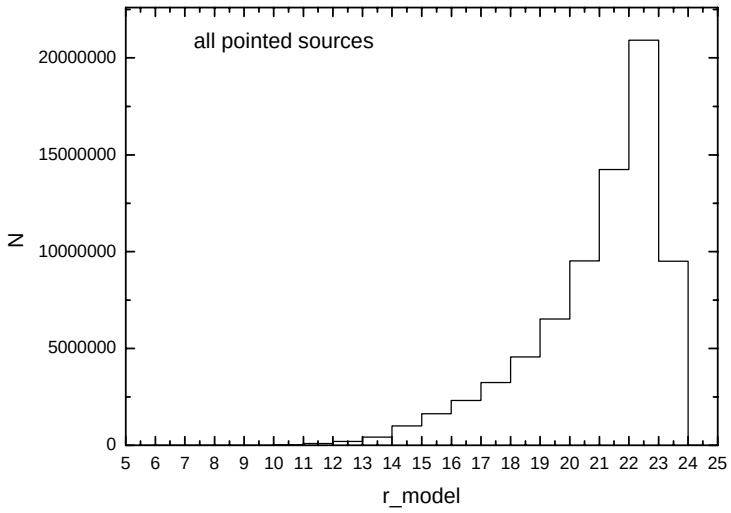


图 5-1 SDSS 巡天中的所有点源随模型 r 星等的统计分布

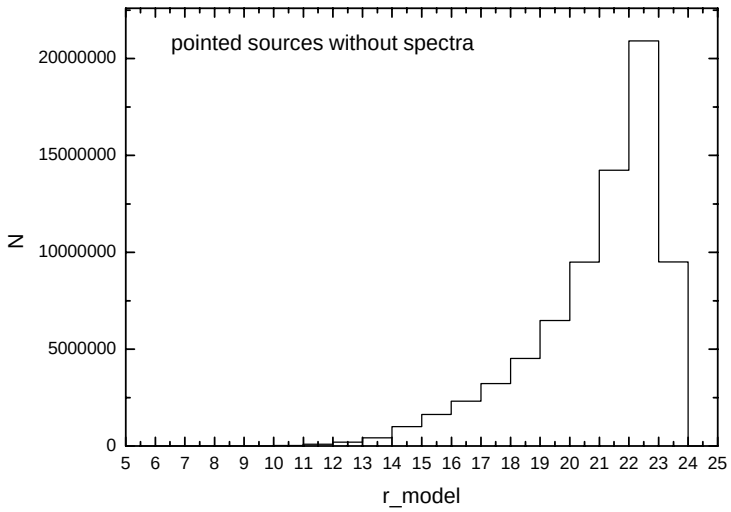


图 5-2 SDSS 巡天中没有光谱观测的点源随模型 r 星等的统计分布

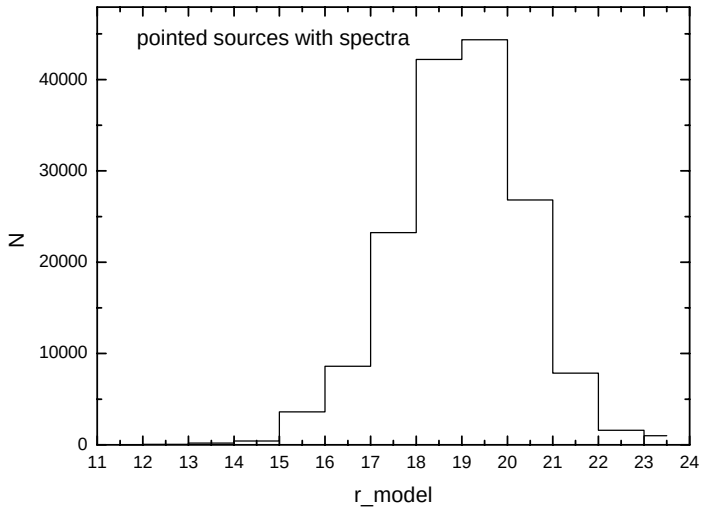


图 5-3 SDSS 巡天中具有光谱观测的点源随模型 r 星等的统计分布

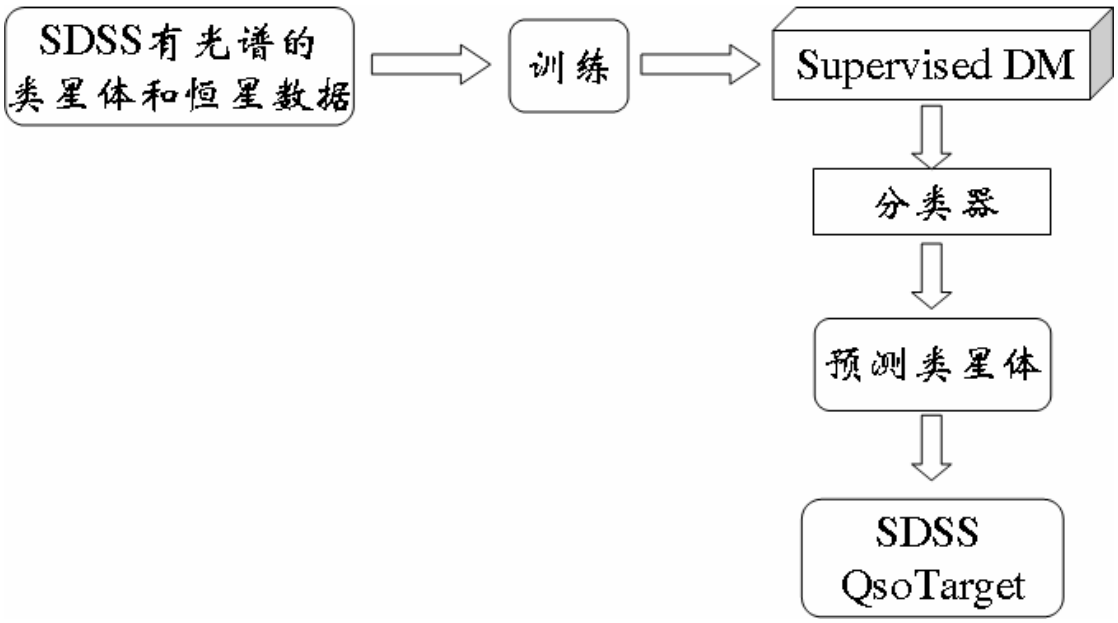


图 5-4 用分类器预测 QsoTarget 数据表

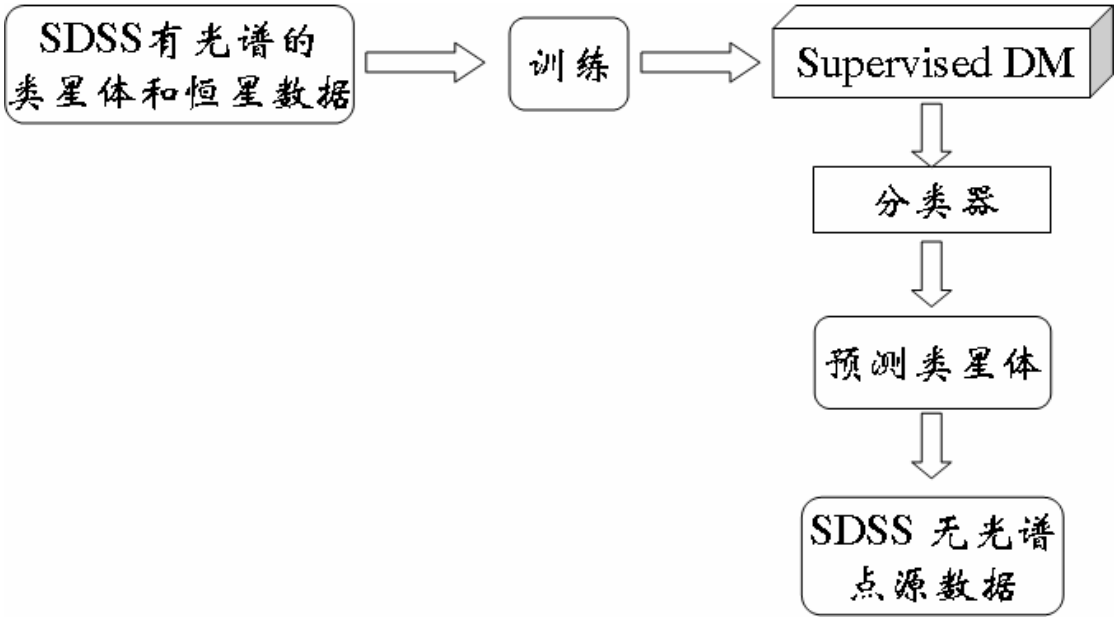


图 5-5 用分类器预测无光谱点源星表

训练数据选择用 SDSS DR2-DR5 的有光谱的恒星和类星体数据（共 185,693 个），参数选择 PSF 星等的四个颜色和 r 星等（u-g、g-r、r-i、i-z、r）。分类器算法选择 kd-tree 算法，参数 N=11。训练后得到分类器。

首先测试的数据为 SDSS DR6 的 QsoTarget 星表，是 SDSS 挑选出来预选类星体候选体的星表，共 321,618 行数据。选取条件为 targetMode=1 且 targetQsotargeted=1，共 170,233 行数据。其中 targetMode=1 是 primary 的源，是指观测了一次的数据；targetQsotargeted=1 是指为认为是类星体的数据。测试结果有 108,228 个类星体，占 63.58%。

SDSS DR6 的 QsoSpec 星表是 QsoTarget 星表中得到光谱证认的类星体星表，是 QsoTarget 星表的子集，共 230,513 行数据，其中 SpecQsoConfirmed=1 即被光谱证认确实是类星体的数据的共 103,647 个，占 44.96%。SDSS DR6 有光谱证认的 103,647 个类星体包括 90,108 个低红移类星体($z < 2.3$)和 13,539 个高红移类星体($z \geq 2.3$) 。

从对 QsoSpec 星表的分析可以看出，符合 SDSS 观测条件的输入星表的类星

体候选体中，被光谱证认确实是类星体的只有 44.96%，这也可以解释我们对 QsoTarget 星表的测试结果（63.58%），因此可以认为分类器的准确率是可靠的。

这里用到的 SDSS 数据是用 SDSS casjob 提供的服务下载的，QsoTarget 星表的下载 SQL 语句如下：

```
Select * from QsoTarget where targetMode=1 and targetQsotargeted=1;
```

随后，我们用分类器来测试 SDSS 无光谱的点源数据。

SDSS 无光谱点源数据的获得过程如下：

首先，从点源星表 Photo（共 74,213,057 行数据）和光谱星表 Spec（共 735,122 行数据）中提取有光谱的点源数据 Photo_spec，共 160,009 行数据，SQL 语句如下：

```
Create table Photo_spec
Select * from Photo, Spec where photo.objId=Spec.objId;
```

然后，从点源星表Photo中除去有光谱的点源数据，得到SDSS无光谱点源数据Photo_nospec，共74,053,096行数据，SQL语句如下：

```
Create table Photo_nospec
select * from Photo
where Photo.objId not in (select Photo_spec.objId from Photo_spec);
```

研究了 QsoTarget 数据基于 i 星等的分布情况。如图 5-6所示，比较了 QsoTarget 数据中所有数据和被分类为恒星的数据基于 i 星等的分布，发现分布几乎一样，高峰值均出现在 19 左右，这说明分类结果没有受到 i 星等的影响。统计了 SDSS 无光谱点源数据基于 i 星等的分布情况，如图 5-7和图 5-8所示，高峰值主要集中在 20 至 21.5 的区域。

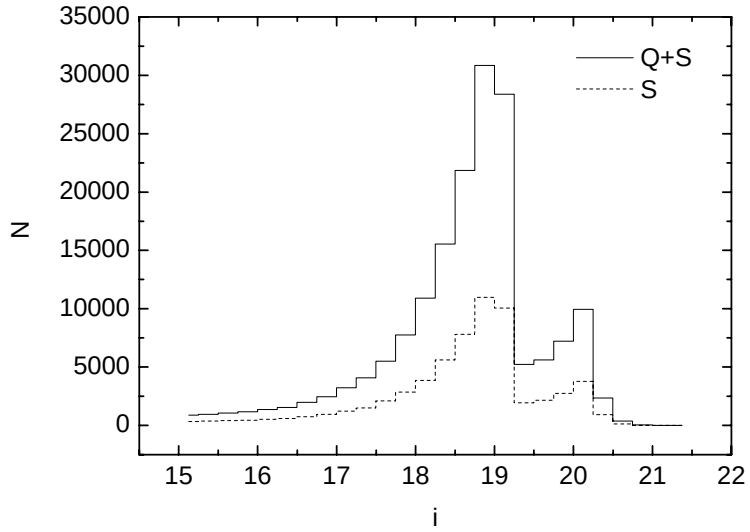


图 5-6 QsoTarget 的全部数据和被分类为恒星的数据 i 星等分布比较

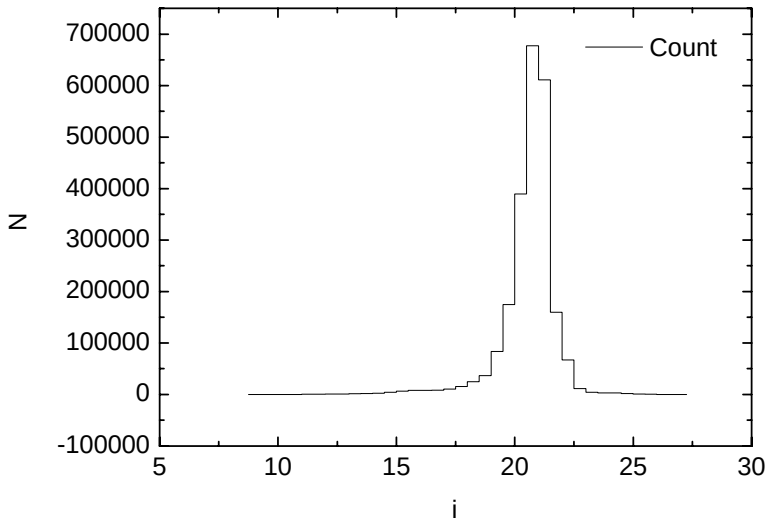


图 5-7 SDSS 无光谱点源数据基于 i 星等分布（1）

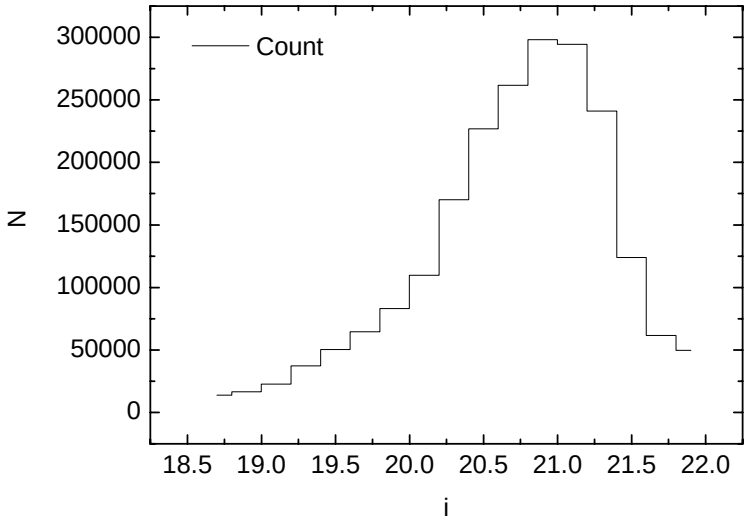


图 5-8 SDSS 无光谱点源数据基于 i 星等分布（2）

基于已知光谱类型的恒星和类星体构建的 kd-tree 分类器，预测了无光谱观测的点源星表，挑选出了类星体候选体，然后针对 LAMOST 的观测条件和星等极限，选出了适合该项目观测的候选体。测试结果预测为类星体的数据共 2,318,275 个，其中按照 LAMOST 的观测条件满足 19.5<i<21、DEC>-10 的共 1,234,683 个。如果更严格点，满足 19.5<i<20.5、DEC>-10，得到的类星体候选体共 560,800 个。因此，我们挑选了符合 LAMOST 观测条件的类星体候选体输入星表共 1,234,683 个数据，这个数据量比 SDSS DR6 的 QsoSpec 的 230,513 个数据要多一百万个。按照观测条件满足 19.1<i<21、DEC>-10，得到的类星体候选体共 1,307,970 个；满足 19.1<i<20.5、DEC>-10，得到的类星体候选体共 634,087 个。这些源的分布如图 5-9至图 5-12所示。图 5-9给出了满足 LAMOST 观测条件的类星体候选体的 i 星等分布（19.5<i<21）。图 5-10描述了满足 LAMOST 观测条件的类星体候选体的 i 星等分布（19.5<i<20.5）。图 5-11给出了满足 LAMOST

观测条件的类星体候选体的 i 星等分布（19.1<i<21）。图 5-12描述了满足 LAMOST 观测条件的类星体候选体的 i 星等分布（19.1<i<20.5）。

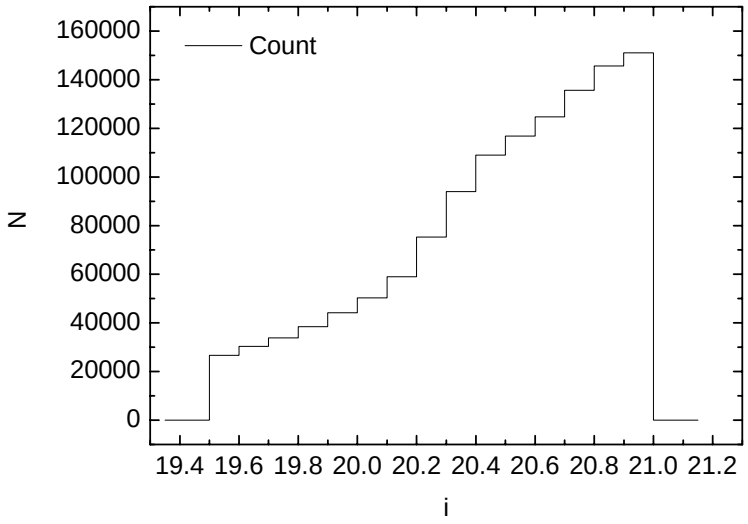


图 5-9 满足 LAMOST 观测条件的类星体候选体的 i 星等分布（19.5<i<21）

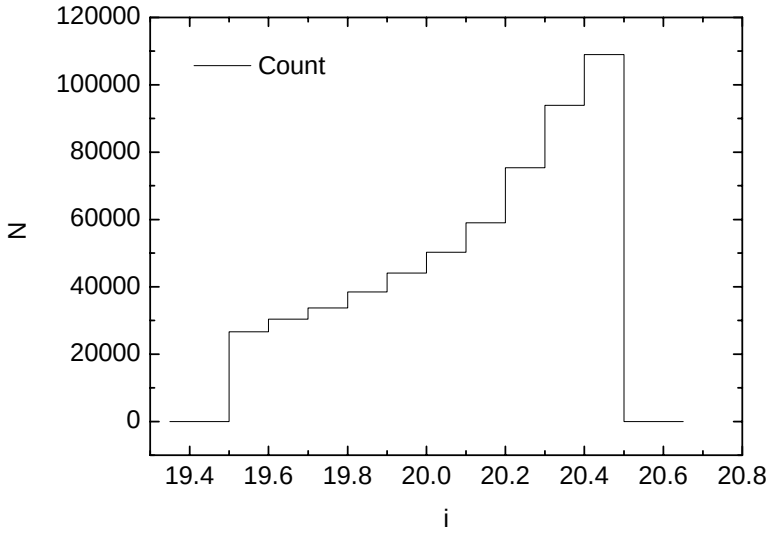


图 5-10 满足 LAMOST 观测条件的类星体候选体的 i 星等分布（19.5<i<20.5）

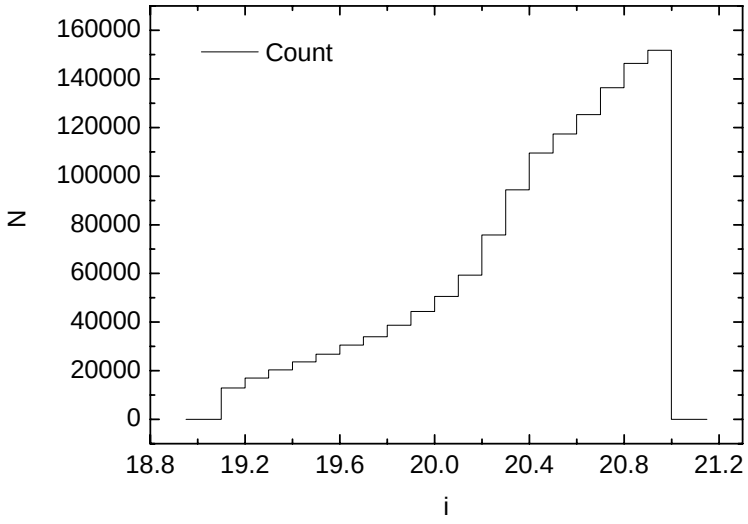


图 5-11 满足 LAMOST 观测条件的类星体候选体的 i 星等分布（19.1<i<21）

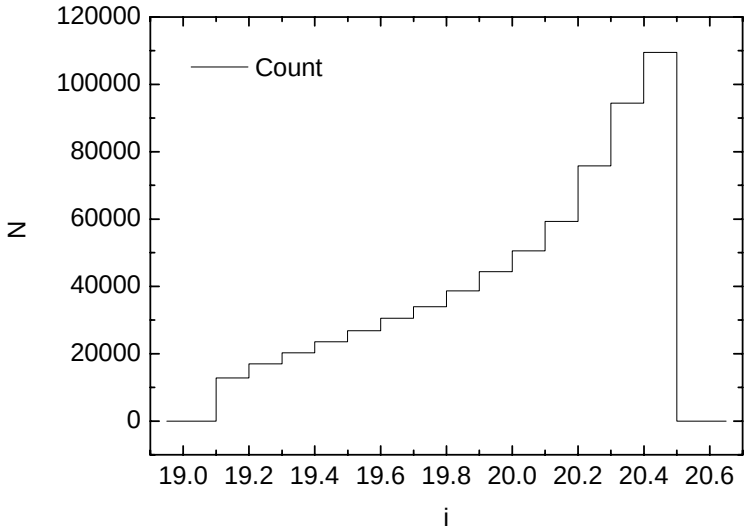


图 5-12 满足 LAMOST 观测条件的类星体候选体的 i 星等分布（19.1<i<20.5）

5.4 小结

本章介绍了 LAMOST 项目的基本情况及观测条件和要求，指出了为 LAMOST 选取类星体候选体的必要性和紧迫性。阐述了各种预选源方法及其成功率，并比较了这些方法的优缺点，通过对比发现数据挖掘方法用于预选源的优越性：高的成功率和适合大样本研究。然后以 SDSS 巡天为例，将最快速、准确、能处理大数据量的 kd-tree 算法应用于为 LAMOST 预选类星体输入星表工作中。首先分析了 SDSS 巡天的类星体输入星表及得到光谱证认的类星体星表，给出了他们的成功率，并与我们用 kd-tree 算法预测的结果相比较，进一步说明该算法的有效性。然后以 SDSS 有光谱证认的恒星和类星体作为训练样本，训练 kd-tree 得到分类器，来预测无光谱证认的点源的类型。从中挑出了满足 LAMOST 观测条件要求的源作为 LAMOST 的类星体候选体的输入样本。

第 6 章 总结和展望

本文主要致力于天文数据融合系统的开发与应用和数据挖掘中的自动分类与回归方法的研究，以及应用自动化分类方法构建分类器，用于为 LAMOST 项目选取类星体候选体。探索了应用 kd-tree、支持矢量机（SVMs）、随机森林（Random Forest）对类星体和恒星及类星体、恒星和星系进行分类。并采用 kd-tree 对 SDSS 巡天的星系的测光红移进行估测。研究了数据挖掘的整个过程，从数据选择、数据清洗、降维去噪、特征选择/抽取、特征加权，按照挖掘任务选取挖掘方法到结果的解释与评估。考虑到 kd-tree 分类时速度快、准确率高的特点，在实际应用方面，用该方法为 LAMOST 挑选出了适合 LAMOST 观测条件的类星体候选体。

首先介绍了基于星表 ReadMe 文件的自动入库工具的开发。有了此工具，用户只需根据模板创建 ReadMe 文件，即可实现星表文件的自动入库。通过该工具能更好地管理和处理海量天文数据，解决了天文数据的海量性和分布性的问题。自动入库工具还能够将星表数据统一进行预处理，如统一赤经赤纬格式，以便使数据融合工具可以对处理过的星表数据进行统一的操作。

接着阐述了海量数据的交叉证认算法。为了提高交叉证认的效率，我们尝试了 B-tree 索引和 HTM 索引交叉证认的算法，并借鉴以上两个算法的优点，引入了 kd-tree 算法找最近邻的思想，首次提出了 HTM 索引分区与 kd-tree 找最近邻算法结合来交叉证认，彻底解决了内存限制和算法复杂度高的问题，并保证了高的证认精度。这里只用到了位置信息，实际上 kd-tree 算法可以选择多种参数来建树，因此除了位置参数外，还可以考虑星表中的其他参数信息，如星等、色指数等。

在自动入库工具和交叉证认算法的研究基础上，实现了一个不依赖于特定数

据库的海量星表数据融合系统（XMaS_VO 系统），其将自动入库、海量数据的交叉证认等工具整合起来批处理地、多任务地进行工作。用户能够方便地使用此服务进行星表的上传及自动入库、交叉证认等工作，也可以自己建立数据库，方便地将此工具移植到任意服务器或个人电脑上。天文学家通过该融合系统可以轻松地进行多波段数据的提取和研究，突破了原来被大数据量束缚而不能开展工作的限制。使用模拟数据的测试证明采用 HTM 索引分区与 kd-tree 找最近邻算法的 XMaS_VO 系统是可靠的、高效的、灵活的，目前已经有许多基于 XMaS_VO 系统的海量星表的交叉证认应用。XMaS_VO 系统减少和缩短了数据准备和预处理过程的工作量和时间，从而方便了海量天文数据挖掘课题的研究。算法研究人员只需关注在算法性能和应用的研究本身，而无需在海量数据处理方面花费太多的时间和精力。

鉴于 XMaS_VO 系统的高效性和可用性，用其交叉证认了多波段星表，如 SDSS、2MASS、USNO、FIRST 和 ROSAT。我们研究了 kd-tree 和 SVMs 算法用光学波段的 SDSS DR5 和红外波段的 2MASS 数据分类类星体和恒星。结果表明 kd-tree 和 SVMs 都是有效的分类器。只考虑准确率，kd-tree 和 SVMs 都对类星体和恒星的分类有很高的准确率，而 kd-tree 创建分类器的速度要比 SVMs 快得多。如果既考虑准确率又考虑速度，kd-tree 是更好的选择。当输入参数越少时，两个算法的性能都较好。而在加入 2MASS 的星等作为参数时，效果反而降低了，可能是因为 J、H、Ks 的极限星等较浅。SDSS 的三种星等中，基于模型星等的结果比基于红化校正模型星等的好，而基于红化校正模型星等的结果又比基于 PSF 星等的要好。当把四个颜色和 r 星等作为输入参数时的性能比五个星等的好，而当输入参数是模型星等的四个颜色和 r 星等（u-g、g-r、r-i、i-z、r）时，得到的结果最好。另外，使用 10 倍的交叉确认方法比测试-训练方法性能要好，但时间开销要大 10 倍之多。这两个算法的分类精度都高达 90%以上，因此可以应用于解决天文学中面临的分类问题，用于多波段天文数据的分类和为大型的巡天项目（比如 LAMOST）预选类星体候选体。

我们还研究了随机森林算法在多波段数据分类中的应用。实验结果表明随机

森林是有效的分类器，用来解决天文学中类星体、恒星和星系的分类问题，尤其是将类星体从恒星中分离出来。分类类星体和恒星时，利用光学和 X 射线波段的数据得到的分类准确率比利用光学和射电波段数据的要高；而分类类星体、恒星和星系时，利用光学和射电波段的数据得到的分类准确率要高。使用参数选择或带权值的参数，算法的性能要比原有参数稍好一些。随机森林算法能评估最重要的分类参数，这一点可被应用于多波段数据的参数抽取和参数加权，来降低分类空间的维数，而且随机森林可以用于探测离群数据。不像人工神经网络算法，随机森林算法不会过度拟合；也不像支持矢量机那样有许多模型参数要调节。随机森林有如此多的优点，故认为其是解决天文学中的分类、参数选择和参数加权、离群数据探测等问题的有效算法。尤其是随机森林算法可以用于探测离群数据，这对想找稀有或未知类型天体或天文现象的天文学家具有十分重要的意义。因此，随机森林算法将会在天文学的研究中发挥出重要的科学价值。

然后，运用 kd-tree 算法来解决天文中参数估计或回归的问题。通过与类似的其他工作相比较发现，kd-tree 具有自身特有的优势，高速快效的查询功能，非常适合大数据量的红移预测，而且其精度与神经网络的结果相当。尽管核回归略优于 kd-tree，可是核回归用于估测红移时不仅速度慢还存在损失点。因此，在将来的工作中可以综合两者的优点，即 kd-tee 的高速性和核回归的高准确性，将两种方法结合使用。从而可以摆脱由天文数据量日益增长的趋势带来的一些算法由于速度太慢而无法实施的困境。

最后，基于 kd-tree 的高准确率和快速性，将其用于选取 LAMOST 的类星体候选体。以有光谱证认的恒星和类星体作为训练集来构造分类器，来预测无光谱证认的点源星表。按照 LAMOST 观测条件满足 19.5<i<21、DEC>-10 的共有 1,234,683 个类星体候选体，满足 19.5<i<20.5、DEC>-10，共有 560,800 个。按照观测条件满足 19.1<i<21、DEC>-10，得到的类星体候选体共 1,307,970 个；满足 19.1<i<20.5、DEC>-10，共有 634,087 个。

综上所述，我们的工作着眼于海量天文数据的存储、融合和挖掘的研究，在研究过程中，解决了包括入库、交叉证认、分类和回归等问题，并成功地将研究

成果应用于包括 LAMOST 预选源等工作中。开发的 XMaS_VO 系统可以方便天文学家轻松地管理数据和获取多波段数据。该系统为虚拟天文台其他工具的开发与发展提供重要的参考和指导，并且促进了相关工具的开发与研究。多种数据挖掘算法的研究与应用丰富了数据挖掘算法在天文学中的应用。纷繁复杂的天文数据（如：海量性、异地异构性、非线性、存在噪声等）对其他算法的发展或现有算法的改进提出了更高的要求。考虑到数据的质量、数量、完备性和高维等特性，面对不同的数据挖掘任务，须慎重准备数据、选择数据挖掘算法及结果的解释与评估。本文中涉及到的对诸多问题的研究和开发的多个工具都存在很多值得进一步发展和改进的地方，比如对海量天文数据融合系统界面的优化、其他数据挖掘算法的研究、并行化数据挖掘的研究以及对面向天文数据特点的数据挖掘工具的开发等。而且可以将这里的算法应用于其他问题的研究，如：星系按形态或哈勃序列的分类、恒星光谱分成不同的次型、恒星物理参数的测定。随着各个虚拟天文台项目的发展和完善，提供的服务和工具将日臻完善，天文学家可以更多、更好、更加便利地从事科学研究，从而推动天文学技术与方法的革新，带动天文学理论的发展。

附录 A　XMaS_VO 各功能模块整合脚本文件

在指定工作目录下，

编辑可执行 Linux Shell 脚本文件 steps_crosswork，然后运行此脚本。

编辑此文件可参考模板文件 steps_crosswork_template

```
./step1 step1_file_putdata
./step2 step2_file_HTMindex
./step3_7 step3_file_crossmath7
./step4 step4_file_sql
./step5 step5_file_getdata
```

编辑 step1_file_putdata 文件可参考模板文件 step1_file_putdata_template

```
GLMC_l350.tbl ReadMe test glmc
```

Step1 是一个可执行文件，内容如下：

```
vi -s script1 $1
sh $1
vi -s script1_end $1
```

script1 文件内容如下：

```
:%g/^$/norm dd
:%s/^/\.\/putdata1 /
:wq
```

script1_end 文件内容如下：

```
:%s/\.\/putdata1 //
:wq
```

putdata1 也是一个可执行文件，内容如下：

```
java -cp /home/gaodan/intodb/ Dbreadme.Putdata
/data0/gaodan/${4}/${1} /data0/gaodan/${4}/${2} $3
```

编辑 **step2_file_HTMindex** 文件可参考模板文件 **step2_file_HTMindex_template**

```
GLIMPSE.GLMC_l350 _glmc 8 GLIMPSE.GLMCl350pix8 RAdeg DEdeg
```

Step2 内容和 **Step1** 内容差不多，内容如下：

```
vi -s script2 $1
sh $1
vi -s script2_end $1
```

script2 和 script2_end 文件内容也差不多，就不再列举。

文件中的 **HTMindex2** 文件内容如下：

```
java -cp /home/gaodan/htm/htmIndex.jar:/home/gaodan/htm
htm.Create_pcode_column $1 $2 $3 $4 $5 $6 id_htm
/data0/gaodan/pix${2}
```

举例说明 Step3 的情况：

编辑 **step3_file_crossmatch1** 可参考模板文件 **step3_file_crossmatch1_template**

```
SDSS_GD.gals TwoMass.twomass_xsc SDSS_GD.sdsspix10_gals
TwoMass.2masspix10_xsc _sdss_gals _2mass ra decl ra decl 0.1
1.5 SDSS_GD.cross_gals_2massxsc twomassmatch
```

可执行文件 **crossmatch3_1** 内容如下：

```
java -cp /home/gaodan/htm/ htm.HTMCross_KDTREE $1 $2 $3 $4
$5 $6 test.distinct_HTM test.id_ra_decl_small
test.id_ra_decl_big id_htm $7 $8 id_htm $9 ${10} ${11} ${12}
${13} ${14} /data0/gaodan/cross${5}${6}1.dat
```

编辑 **step4_file_sqldata** 文件可参考模板文件 **step4_file_sqldata_template**

```
gaodan.cross_glmcl350_2mass_data GLIMPSE.GLMC_l350
TwoMass.twomass_psc gaodan.cross_glmcl350_2mass _glmc
_2mass "t1.designation,t2.ra,t2.decl,t3.d"
```

Step4 内容有所变化，多了一步操作数据库的语句如下：

```
rm step4_temporary
```

```
vi -s script4 $1
sh $1
vi -s script4_end $1
/usr/bin/mysql –p 数据库密码 <step4_temporary
rm step4_temporary
```

可执行文件 sqldata4 将生成的 SQL 语句存入临时文件 step4_temporary，如下：

```
echo "Create table ${1} select ${7} from $2 as t1, $3 as t2 ,
$4 as t3 where t3.id${5}=t1.id_htm and t3.id${6}=t2.id_htm;"
>>step4_temporary
```

编辑 step5_file_getdata 文件可参考模板文件 step5_file_getdata_template

```
gaodan cross_glmcl350_2mass_data /cross/glmc/
```

可执行文件 getdata5 内容如下：

```
java -cp /home/gaodan/htm/ htm.Testrs ${1}.${2}
lftp -e "open 192.168.3.19;put -E /data0/gaodan/${2}.dat -o
${3}${2}.dat;exit"
```

附录 B　XMaS_VO 使用详细说明

数据库的表说明：

数据表：如 GLIMPSE.GLMC_l350，其中数据库名由 step1 指定，表名为数据文件名。

HTM 索引表：如 GLIMPSE.GLMC_l350pix8，其中数据库名和数据表一样，也可另行指定，表名在数据表名基础上加上“pix”及 HTM 索引的级数。

交叉证认表：如 gaodan.cross_glmcl350_2mass，数据库名自定，表名为 cross_小表简称_大表简称，表中只有交叉证认两表的赤径、赤纬、id_htm 主健以及角距离 d 的信息。

交叉证认提取参数表：如 gaodan.cross_glmcl350_2mass_data，表名比交叉证认表多了一个“_data”后缀，也可生成并提取出任意名字的交叉证认提取参数表。

参数说明：

HTM 后缀：HTM 表中列的后缀名

赤径：赤径列名（一般为 RAdeg）

赤纬：赤纬列名（一般为 DEdeg）

误差：误差半径（arcsec）

生成的临时文件：（方便定期删除）

所有生成的临时文件都在/data0/gaodan/目录下。

Step1 /data0/gaodan/data_随机数.dat

Step2 /data0/gaodan/pix_HTM 后缀_随机数.dat

Step3 /data0/gaodan/cross_小表 HTM 后缀_大表 HTM 后缀 1.dat

Step5 /data0/gaodan/交叉证认提取参数表表名.dat

生成的临时数据库：（执行结束自动删除）

Test.distinct_HTM

Test.id_ra_decl_small

Test.id_ra_decl_big

Step1 自动入库：

功能：用户输入星表 ReadMe 文件、数据文件名、入库的数据库名及星表所在的文件包名，程序自动建表、入库、将时分秒形式转换成度（RAdeg、DEdeg）、建索引、建 id_htm 主键。

注：星表 ReadMe 文件、数据文件需放在/data0/gaodan/文件包/路径下。

编辑 step1_file_putdata，参考模板 step1_file_putdata_template

参数 4 个：星表数据文件名 ReadMe 文件名 数据库名 文件包名

如：`GLMC_l350.tbl ReadMe GLIMPSE glmc`

Step2 建 HTM 索引：

功能：程序根据数据表的坐标数据（单位：度；历元：J2000）计算出对应 HTM 索引的 pcode 值，为星表建立 HTM 索引，将 id_htm 主键和 pcode 值两列新建 HTM 索引表。

注：误差半径为 5arcsec 选 8level，为 30arcsec 选 6level, 小于 5arcsec 的都选 8level。数据表主键必须为 id_htm。

编辑 step2_file_HTMindex，参考模板 step2_file_HTMindex_template

参数 6 个：数据表、HTM 后缀、HTM 级数、HTM 索引表、星表赤径、星表赤纬

如：GLIMPSE.GLMC_l350 _glmc 8 GLIMPSE.GLMCl350pix8 RAdeg DEdeg

Step3 交叉证认：

功能：程序根据两个数据表及其 HTM 索引表、误差半径进行交叉证认，得到交叉证认表，并分别给交叉证认表中的两表 id 列建索引，方便进一步的数据参数提取。

根据需要选择相应的程序 step3_file_crossmatch（1，2，3，6，7，8，9，10）

注：交叉证认的两表主键必须为 id_htm。

1）当用户要进行交叉证认的两个星表 HTM 级数不同，而用户需要 1 对 N 的交叉证认结果的时候。（当 N=1 时即为 1 对最近邻）。

编辑 step3_file_crossmatch6，参考模板 step3_file_crossmatch6_template

参数 14 个：小表数据表、大表数据表、小表 HTM 索引表、大表 HTM 索引表、小表 HTM 后缀、大表 HTM 后缀、小表赤径、小表赤纬、大表赤径、大表、小表误差、大表误差、交叉证认表、N

如：IRAS.main TwoMass.twomass_psc gaodan.iraspix6 TwoMass.2masspix10 _iras _2mass RAdeg DEdeg ra decl 20 0.1 gaodan.cross_iras_2mass 100

2）当用户要进行交叉证认的两个星表 HTM 级数不同，而用户需要 1 对 1 交叉证认结果，当 1 对 2 或更多则删除证认结果的时候。

编辑 step3_file_crossmatch7，参考模板 step3_file_crossmatch7_template

参数 13 个：小表数据表、大表数据表、小表 HTM 索引表、大表 HTM 索引表、小表 HTM 后缀、大表 HTM 后缀、小表赤径、小表赤纬、大表赤径、大表赤纬、小表误差、大表误差、交叉证认表

如：GLIMPSE.GLMC_l350 TwoMass.twomass_psc GLIMPSE.GLMCl350pix8 TwoMass.2masspix10 _glmc _2mass RAdeg DEdeg ra decl 3 0.1 gaodan.cross_glmcl350_2mass

3）当用户要进行交叉证认的两个星表具有相同的 HTM 级数，而用户需要 1 对 N 的交叉证认结果的时候。（当 N=1 时即为 1 对最近邻）。。

编辑 step3_file_crossmatch3，参考模板 step3_file_crossmatch3_template

参数 14 个：小表数据表、大表数据表、小表 HTM 索引表、大表 HTM 索引表、小表 HTM 后缀、大表 HTM 后缀、小表赤径、小表赤纬、大表赤径、大表赤纬、小表误差、大表误差、交叉证认表、N

```
如 ： first.first TwoMass.twomass_psc gaodan.firstpix10 TwoMass.2masspix10 _first _2mass RAdeg DEdeg ra decl 5 0.1 gaodan.cross_first_2mass 1
```

4）当用户要进行交叉证认的两个星表 HTM 级数相同，而用户需要 1 对 1 交叉证认结果，当 1 对 2 或更多则删除证认（即为 1 对 1）结果的时候。

编辑 step3_file_crossmatch8，参考模板 step3_file_crossmatch8_template

参数 13 个：小表数据表、大表数据表、小表 HTM 索引表、大表 HTM 索引表、小表 HTM 后缀、大表 HTM 后缀、小表赤径、小表赤纬、大表赤径、大表赤纬、小表误差、大表误差、交叉证认表

```
如 ： first.first TwoMass.twomass_psc gaodan.firstpix10 TwoMass.2masspix10 _first _2mass RAdeg DEdeg ra decl 5 0.1 gaodan.cross_first_2mass
```

5）当用户要进行交叉证认的两个星表具备相同 HTM 级数，用户要进行 1 对最近邻的无损耗证认，证认结果中有一列 boolean 类型的 match column，证认结果如有证认源，此列值为 1，否则为 0。

编辑 step3_file_crossmatch1，参考模板 step3_file_crossmatch1_template

参数 14 个：小表数据表、大表数据表、小表 HTM 索引表、大表 HTM 索引表、小表 HTM 后缀、大表 HTM 后缀、小表赤径、小表赤纬、大表赤径、大表赤纬、小表误差、大表误差、交叉证认表、match 列名

```
如：SDSS_GD.gals TwoMass.twomass_xsc SDSS_GD.sdsspix10_gals TwoMass.2masspix10_xsc _sdss_gals _2mass ra decl ra decl 0.1 1.5 SDSS_GD.cross_gals_2massxsc twomassmatch
```

6）当误差半径较小时，用户给出一个误差半径值 n，而要进行交叉证认的两个星表具备相同 HTM 级数，用户要进行的是 1 对最近邻的无损耗证认。证认结果中有一列 boolean 类型的 match column，证认结果如有证认源，此列值为 1，否则为 0。

编辑 step3_file_crossmatch2，参考模板 step3_file_crossmatch2_template

参数 13 个：小表数据表、大表数据表、小表 HTM 索引表、大表 HTM 索引表、小表 HTM 后缀、大表 HTM 后缀、小表赤径、小表赤纬、大表赤径、大表赤纬、误差、交叉证认表、match 列名

如：`SDSS_GD.gals TwoMass.twomass_xsc SDSS_GD.sdsspix10_gals TwoMass.2masspix10_xsc _sdss_gals _2mass ra decl ra decl 2 SDSS_GD.cross_gals_2massxsc twomassmatch`

7）当误差半径较小时，用户给出一个误差半径值 n，而要进行交叉证认的两个星表具备相同 HTM 级数。用户要进行 1 对 1 的交叉证认。

编辑 step3_file_crossmatch9，参考模板 step3_file_crossmatch9_template

参数 12 个：小表数据表、大表数据表、小表 HTM 索引表、大表 HTM 索引表、小表 HTM 后缀、大表 HTM 后缀、小表赤径、小表赤纬、大表赤径、大表赤纬、误差、交叉证认表

如：`SDSS_GD.gals TwoMass.twomass_xsc SDSS_GD.sdsspix10_gals TwoMass.2masspix10_xsc _sdss_gals _2mass ra decl ra decl 2 SDSS_GD.cross_gals_2massxsc`

8）当误差半径较小时，用户给出一个误差半径值 n，而要进行交叉证认的两个星表具备相同 HTM 级数。用户要进行 1 对最近邻的交叉证认。

编辑 step3_file_crossmatch10，参考模板 step3_file_crossmatch10_template

参数 13 个：小表数据表、大表数据表、小表 HTM 索引表、大表 HTM 索引表、小表 HTM 后缀、大表 HTM 后缀、小表赤径、小表赤纬、大表赤径、大表赤纬、误差、交叉证认表、N

如：SDSS_GD.gals TwoMass.twomass_xsc SDSS_GD.sdsspix10_gals TwoMass.2masspix10_xsc _sdss_gals _2mass ra decl ra decl 2 SDSS_GD.cross_gals_2massxsc 1

9）当用户要进行交叉证认的两个星表具有相同的 HTM 级数，而用户分别需要 1 对 N 和 1 对 1 的交叉证认结果的时候。（N 的值没有限制）。

编辑参数文件 step3_file_crossmatch11，参考模板 step3_file_crossmatch11_template

参数 13 个：小表数据表、大表数据表、小表 HTM 索引表、大表 HTM 索引表、小表 HTM 后缀、大表 HTM 后缀、小表赤径、小表赤纬、大表赤径、大表赤纬、小表误差、大表误差、交叉证认表。

如：first.first TwoMass.twomass_psc gaodan.firstpix10 TwoMass.2masspix10 _first _2mass RAdeg DEdeg ra decl 5 0.1 gaodan.cross_first_2mass

10) 当用户要进行交叉证认的两个星表 HTM 级数不同，而用户分别需要 1 对 N 和 1 对 1 的交叉证认结果的时候。（N 的值没有限制）。

编辑参数文件 step3_file_crossmatch12，参考模板 step3_file_crossmatch12_template

参数 13 个：小表数据表、大表数据表、小表 HTM 索引表、大表 HTM 索引表、小表 HTM 后缀、大表 HTM 后缀、小表赤径、小表赤纬、大表赤径、大表、小表误差、大表误差、交叉证认表。

如：IRAS.main TwoMass.twomass_psc gaodan.iraspix6 TwoMass.2masspix10 _iras _2mass RAdeg DEdeg ra decl 20 0.1 gaodan.cross_iras_2mass

Step4 提取交叉证认星表参数，生成交叉证认提取参数表

编辑 step4_file_sqldata，参考模板 step4_file_sqldata_template

参数 7 个：交叉证认提取参数表、小表数据表、大表数据表、交叉证认表、小表 HTM 后缀、大表 HTM 后缀、提取的参数（包括 t1 表中提取的参数、t2 表中提取的参数、t3 表的角距离参数 d）。

如：gaodan.cross_glmcl350_2mass_data GLIMPSE.GLMC_l350 TwoMass.twomass_psc gaodan.cross_glmcl350_2mass _glmc _2mass “t1.designation as designation_glmc,t2.ra,t2.decl,t3.d”

若对提取所有表的参数的情况，

如：gaodan.cross_glmcl350_2mass_data GLIMPSE.GLMC_l350 TwoMass.twomass_psc gaodan.cross_glmcl350_2mass _glmc _2mass “*”

注：参数中有空格的需把整个参数打上引号，如上例。

step5 从数据库中取出数据并 lftp 传到本机

编辑 step5_file_getdata，参考模板 step5_file_getdata_template

参数 3 个：交叉证认提取参数表数据库名、交叉证认提取参数表表名、文件下载路径

如：gaodan cross_glmcl350_2mass_data /cross/glmc/

生成目标文件：D:/share/文件下载路径/交叉证认提取参数表表名.dat。

注：数据库名、表名要分开写，文件下载路径要先建立。

Lftp 192.168.3.19　根目录 D:/share/。

附录 C　XMaS_VO 移植方法

要将 XMaS_VO 移植到别的机器上，只需要简单的修改下面的几个属性即可。

1　修改文件放置路径：/data0/gaodan/，如改成/ns1/pipeline/

2　修改 XmaS_VO 程序放置路径：/home/gaodan/steps_tar/，如改成/ns1/pipeline/bin/。在程序放置路径下，将工具压缩成 xmatch.tar（压缩命令如下所示），并在新机器的新路径下解压。

```
Tar –cf xmatch.tar Dbreadme/ edu/ htm/ htmIndex.jar steps/
```

3　修改数据库配置文件：Dbreadme 包和 htm 包中的 org 包

4　修改 Step5 中 lftp 的 ip 地址

5　数据库中必须有 Test 数据库

6　数据表中必须有 id_htm 主键列，是 int 类型的逐一递增整数

附录 D　ExecPlan 详细设计

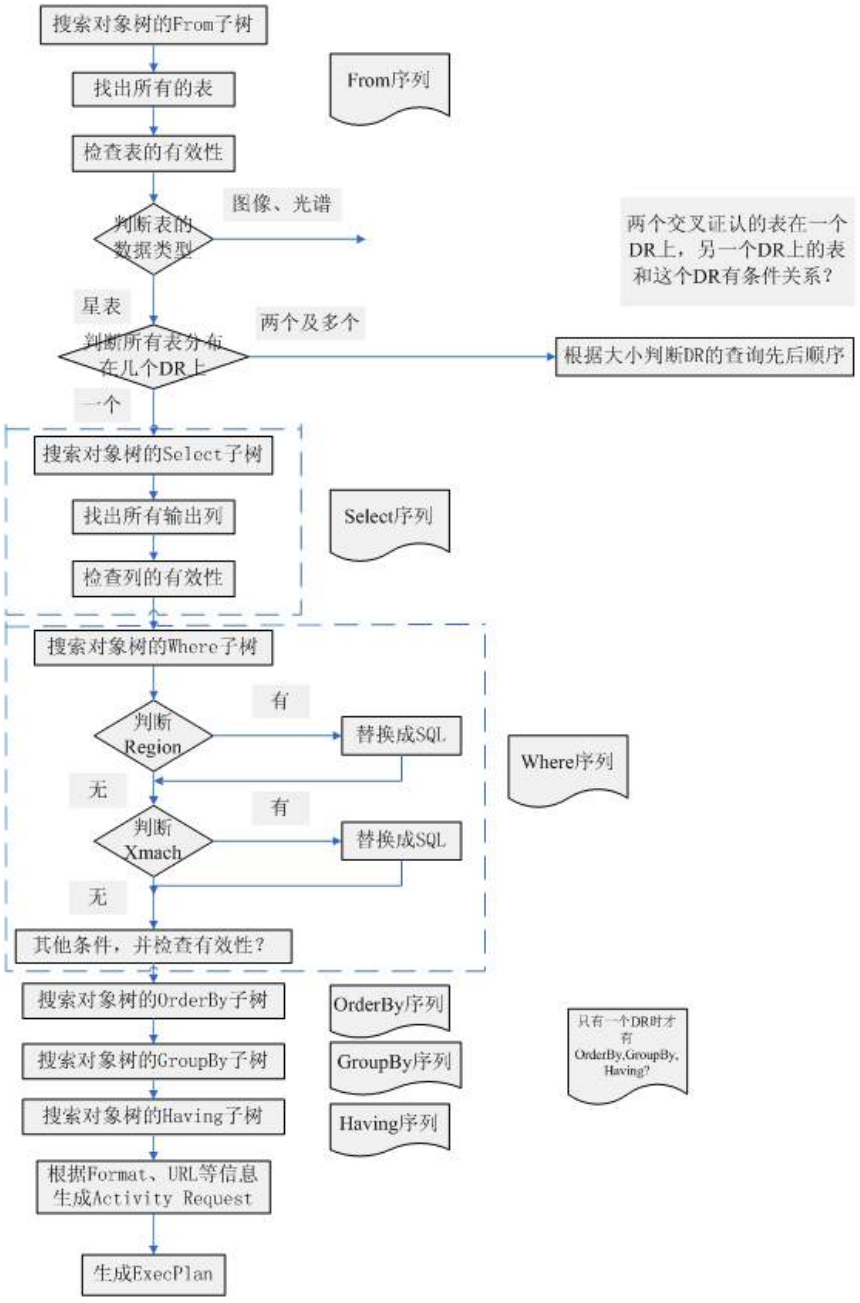


图 D 1　ExecPlan 执行计划的流程图（1）

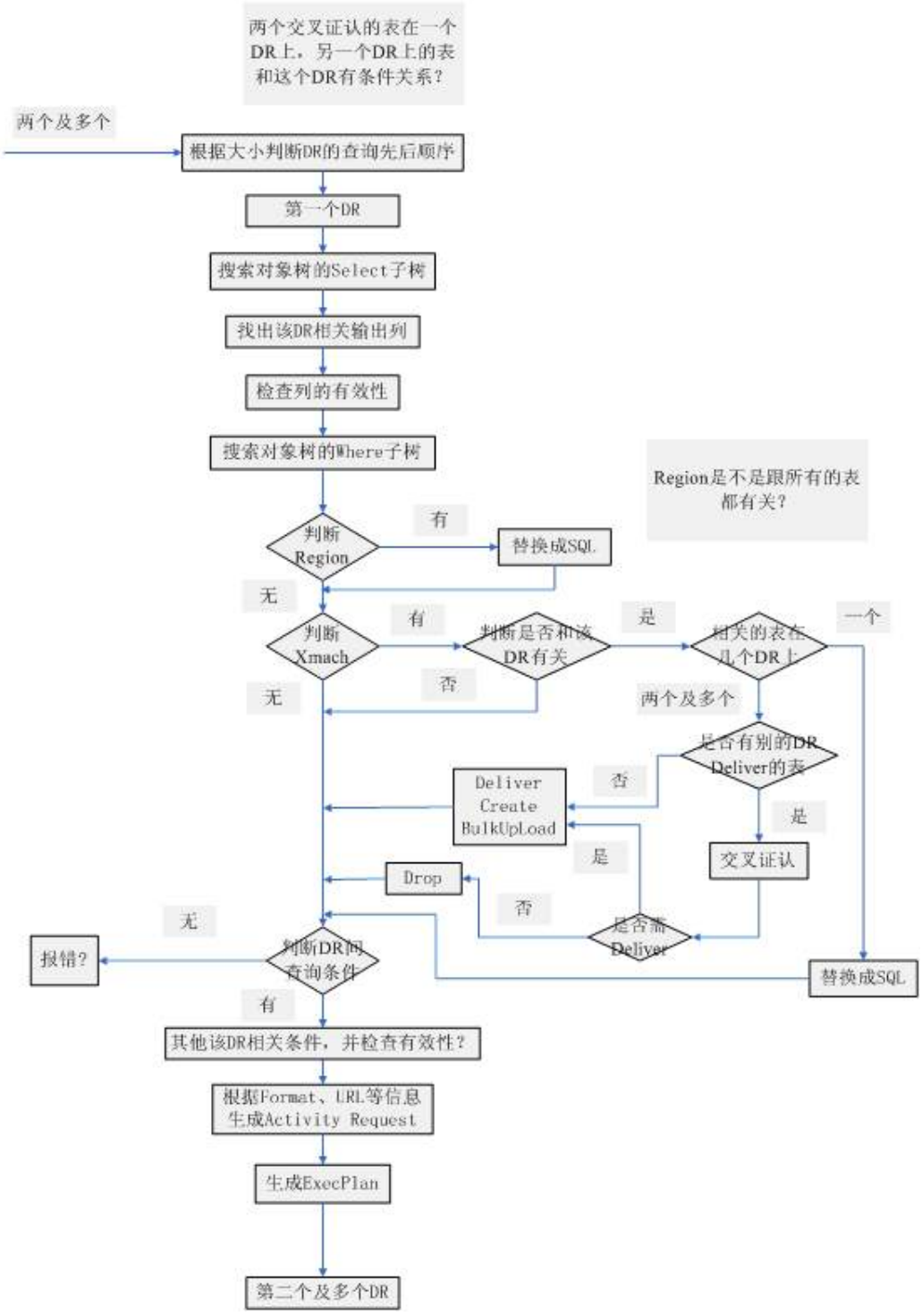


图 D 2 ExecPlan 执行计划的流程图（2）

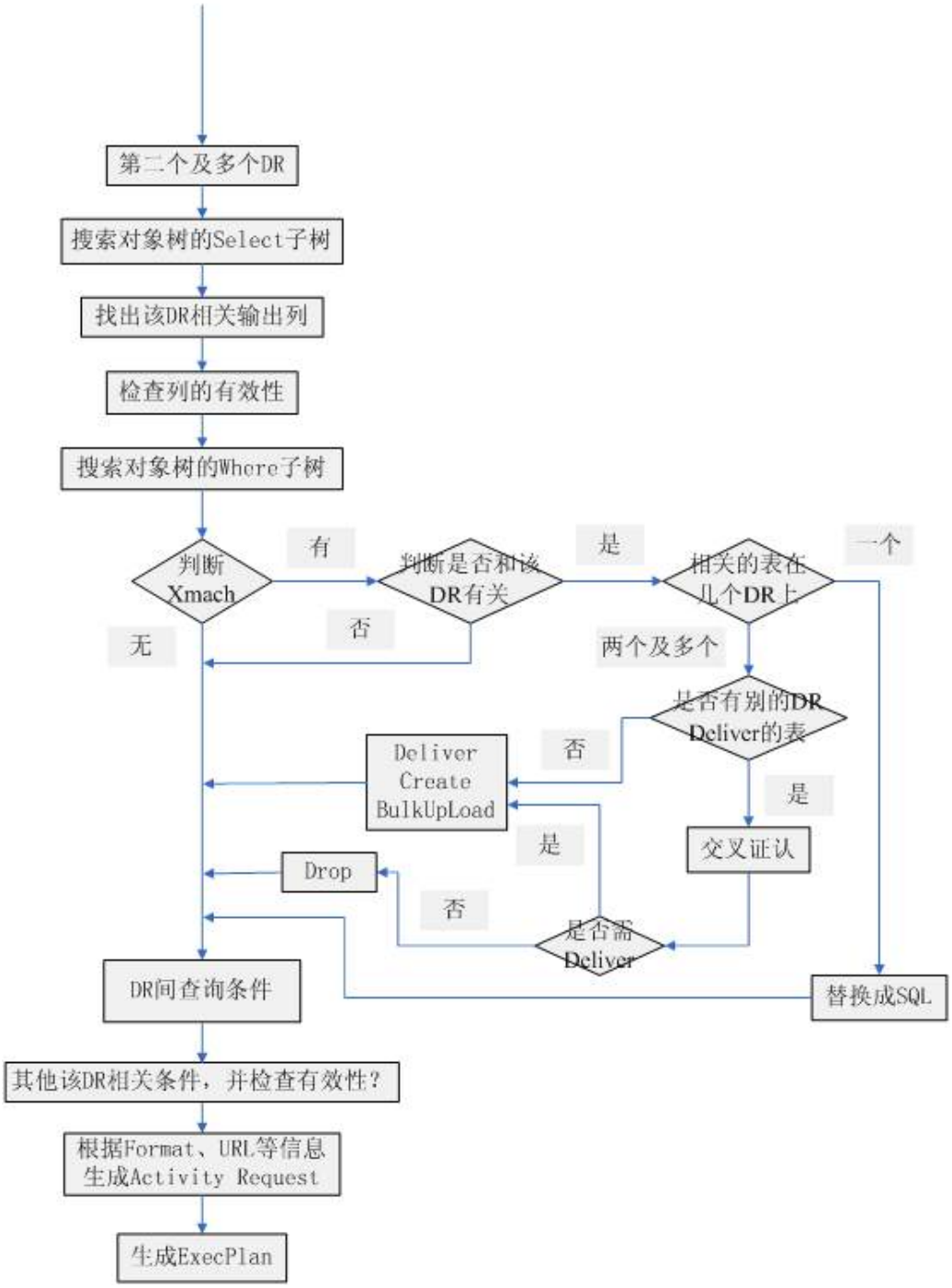


图 D 3　ExecPlan 执行计划的流程图（3）

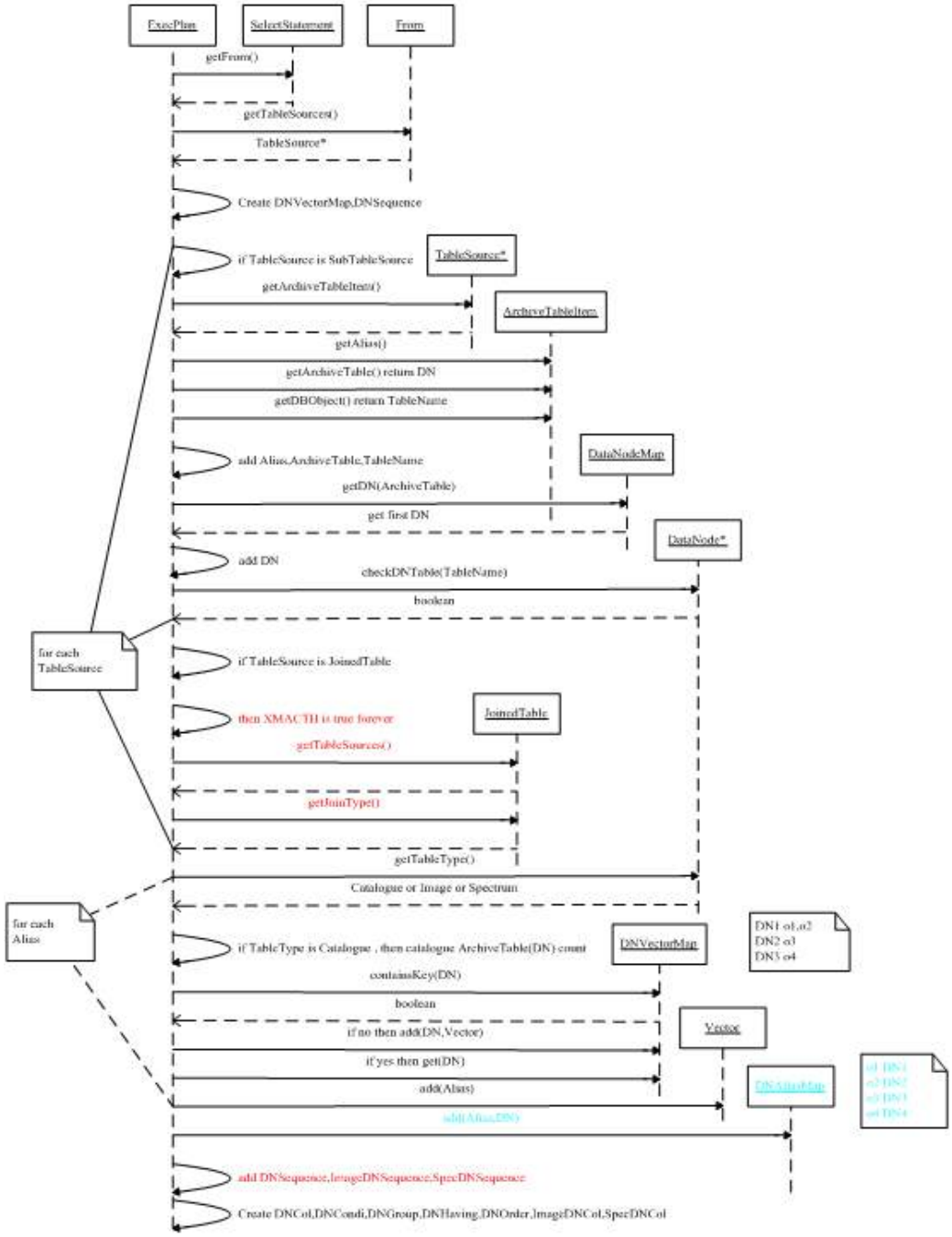


图 D 4 ExecPlan 执行计划 From 语句序列图

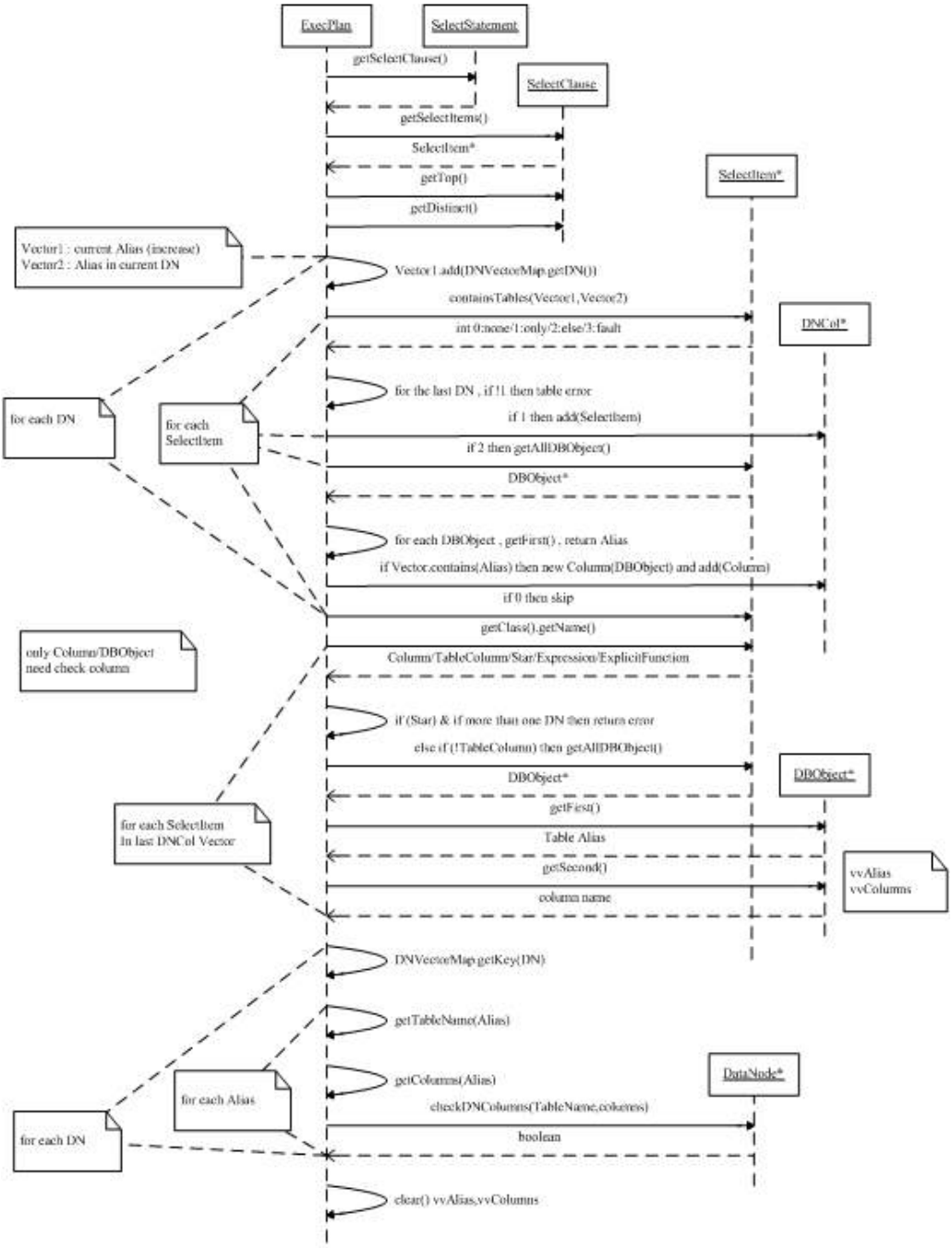


图 D 5　ExecPlan 执行计划 Select 语句序列图

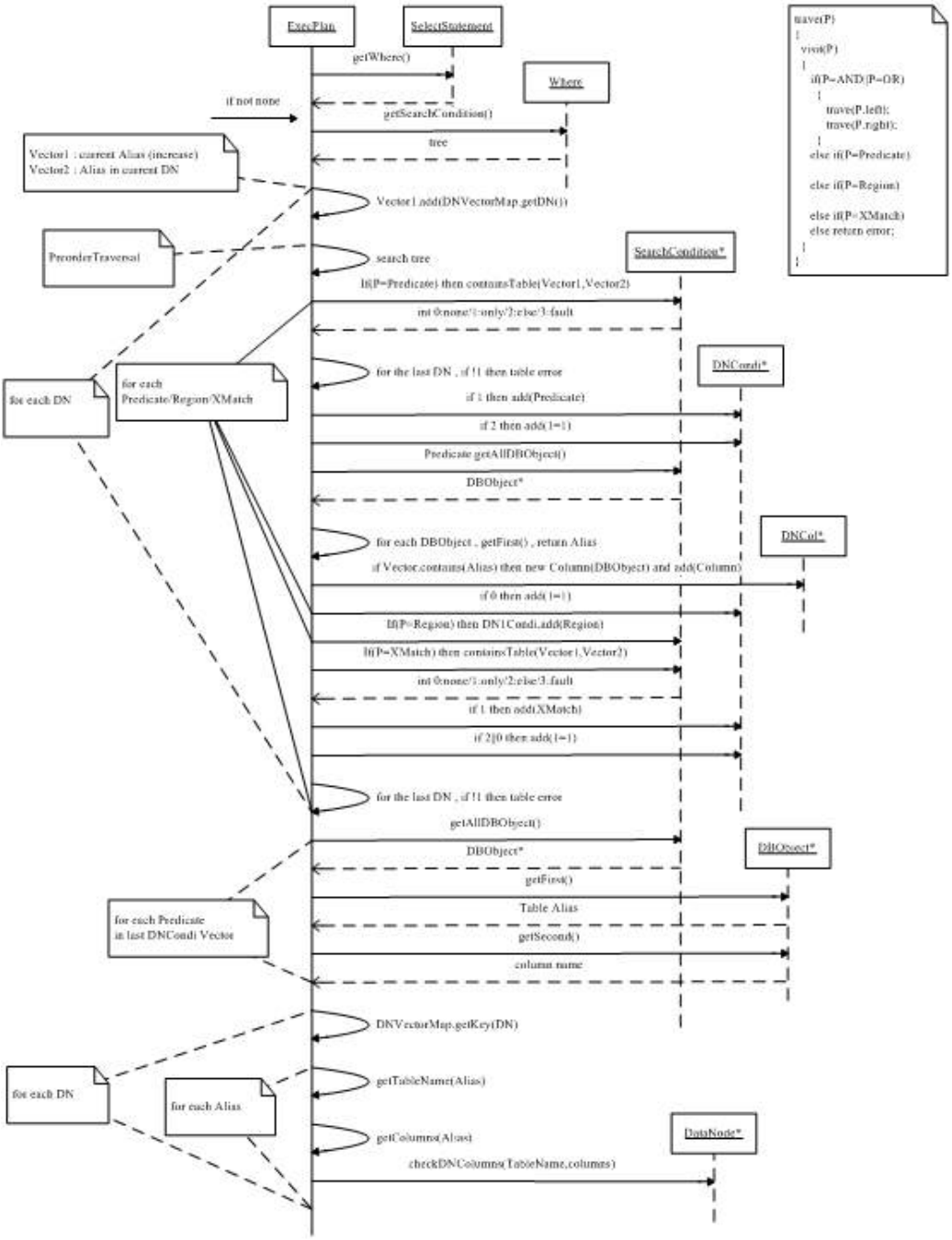


图 D 6 ExecPlan 执行计划 Where 语句序列图

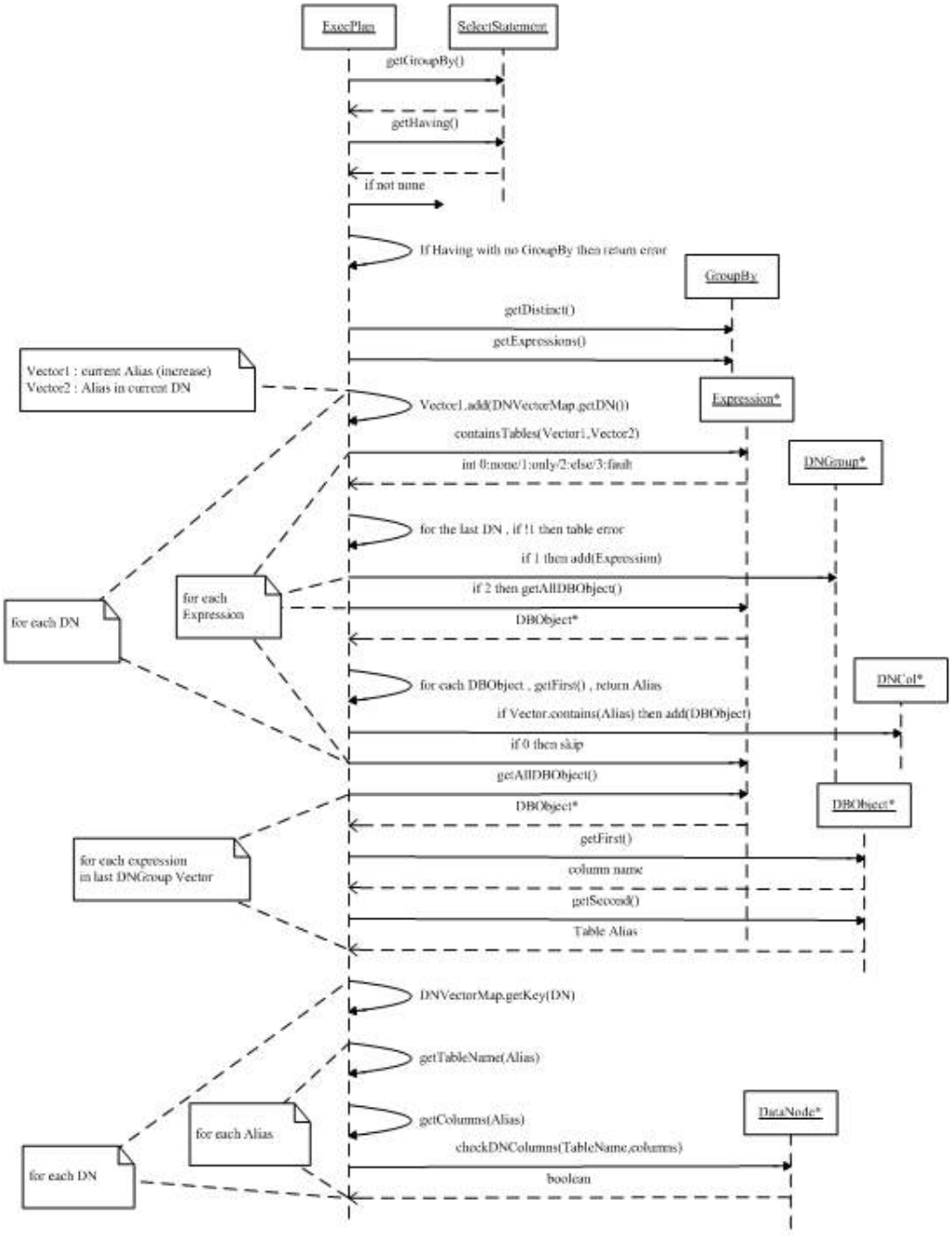


图 D 7　ExecPlan 执行计划 GroupBy、Having 语句序列图

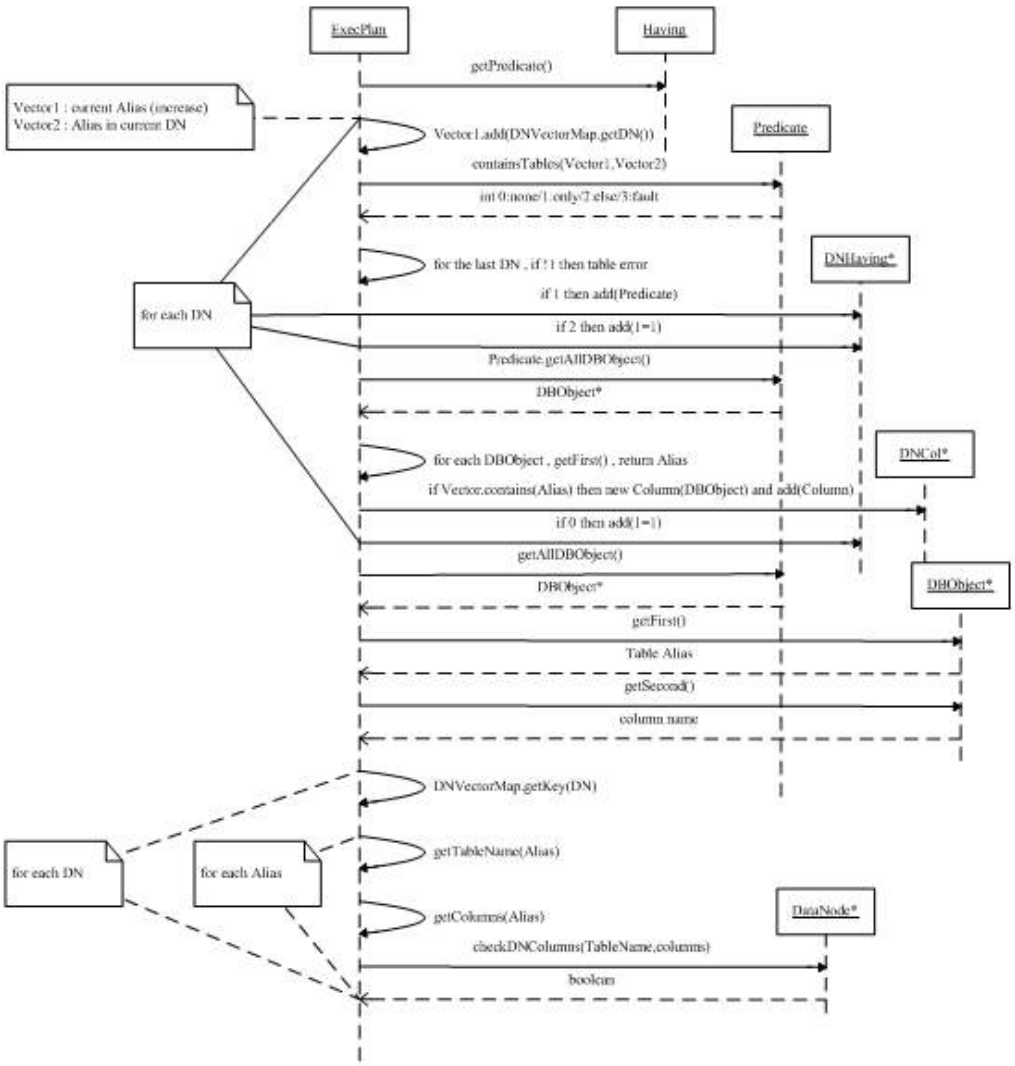


图 D 8 ExecPlan 执行计划 Having 语句序列图

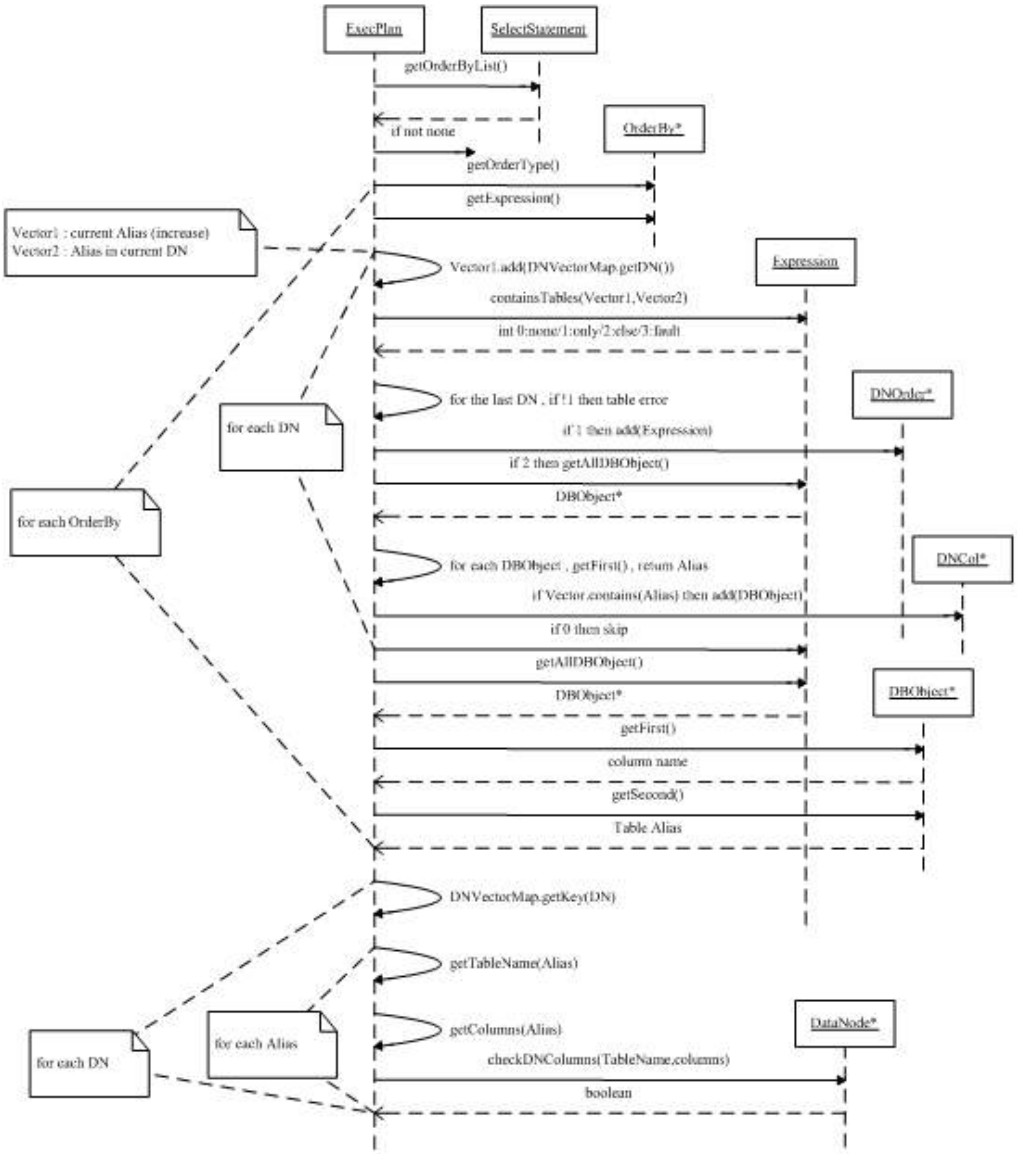


图 D 9　ExecPlan 执行计划 OrderBy 语句序列图

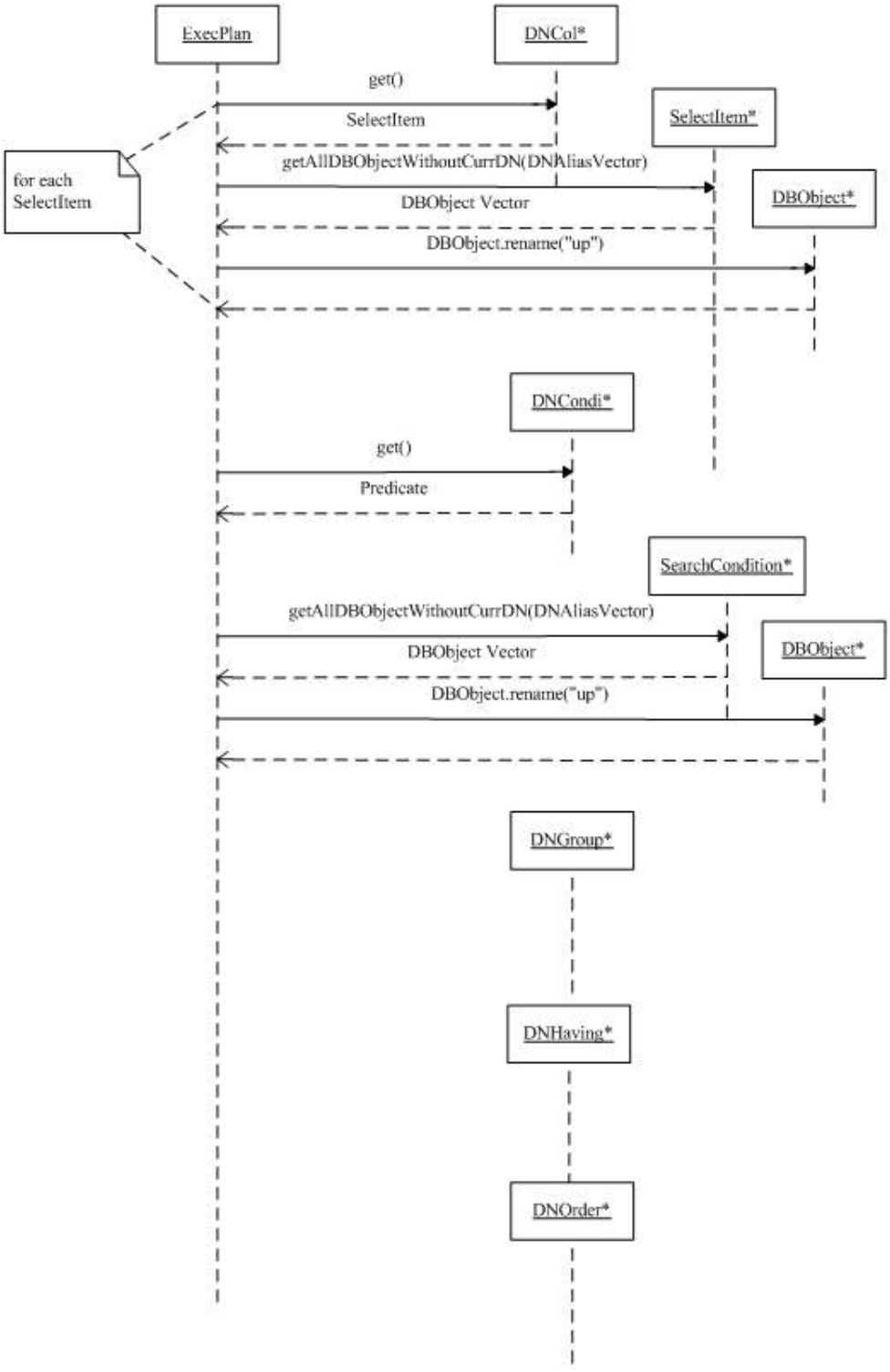


图 D 10 ExecPlan 执行计划参数名替换序列图

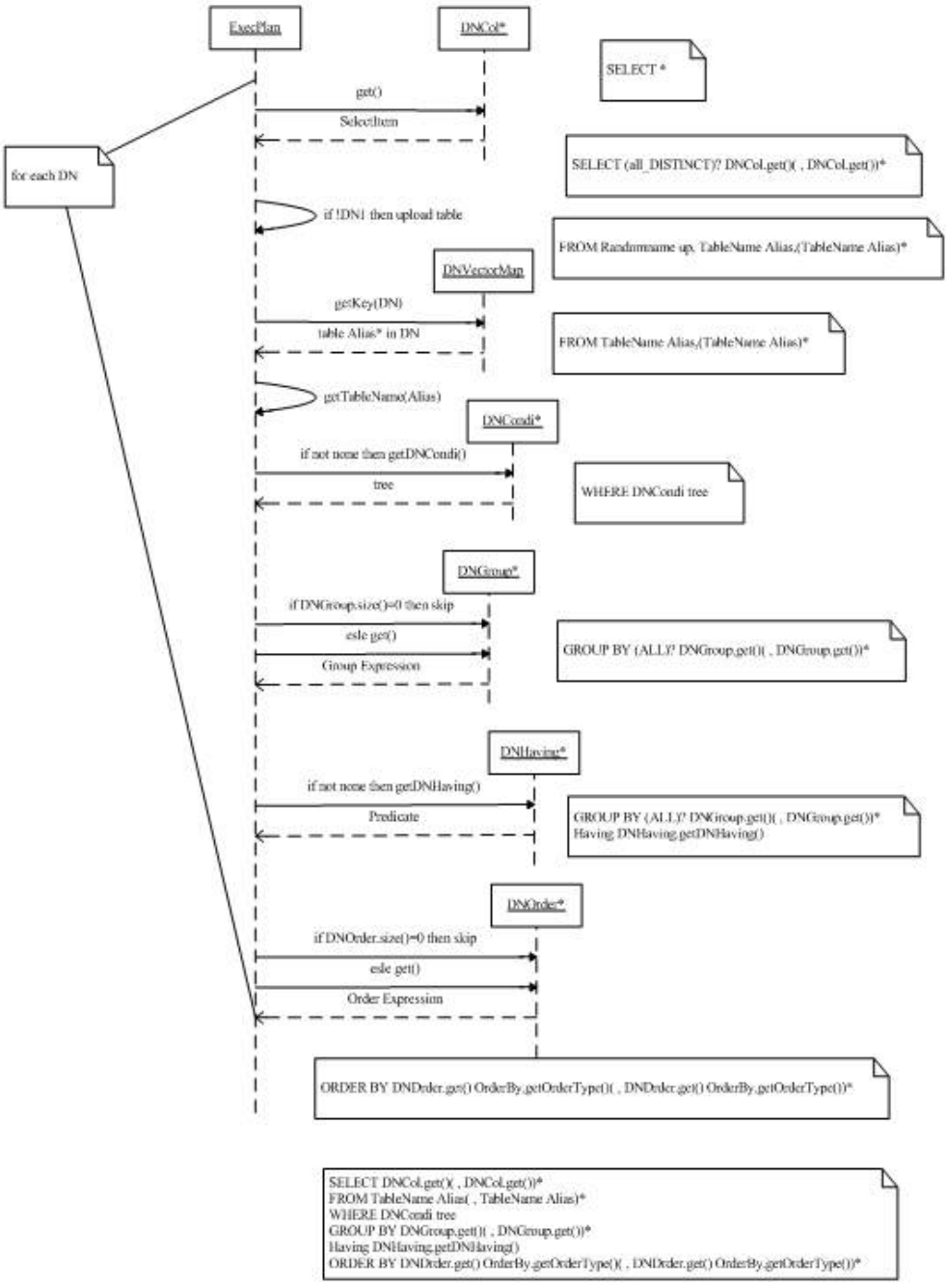


图 D 11　ExecPlan 执行计划生成 ActivityRequest 序列图

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] CDS, http://cdsweb.u-strasbg.fr/.
[3] VizieR, http://vizier.u-strasbg.fr/.
[4] Aladin, http://aladin.u-strasbg.fr/.
[5] OpenSkyQuery, http://openskyquery.net/.
[6] ADQL, http://www.ivoa.net/Documents/latest/ADQL.html/.
[7] Mirage, http://cm.bell-labs.com/who/tkh/mirage/index.html/.
[8] Szalay A,Gray J. The World-Wide Telescope. Science, 2001, 293: 203.
[9] 赵永恒. 互联网时代的天文学革命: 虚拟天文台. 科学, 2002, 54(2): 13.
[10] VOStat, http://www.vostat.org/.
[11] VOPlot, http://vo.iucaa.ernet.in/~voi/voplot.htm/.
[12] Jim Gray, Alex Szalay. Where the Rubber Meets the Sky: Bridging the Gap between Databases and Science. MSR-TR-2004-110.
[13] 张彦霞,赵永恒,崔辰州. 天文学中的数据挖掘和知识发现. 天文学进展, 2002, 20(4): 312.
[14] LAMOST, http://www.lamost.org/.
[15] BADC, 北京天文数据中心, http://badc.lamost.org/.
[16] China-VO, http://www.china-vo.org/.
[17] 崔辰州, 虚拟天文台系统设计. [博士学位论文], 中国科学院研究生院, 2003.
[18] FITS, http://fits.gsfc.nasa.gov/.
[19] IVOA, http://www.ivoa.org/.
[20] Resource Metadata for the Virtual Observatory, http://www.ivoa.net/Documents/latest/RM.html/.
[21] 星表搜索服务, http://badc.lamost.org/cnweb/modules/catsquery/.
[22] CFITSIO, http:// heasarc.gsfc.nasa.gov/docs/software/fitsio/fitsio.html/.
[23] Simbad, http://simbad.u-strasbg.fr/.
[24] MAST, http://archive.stsci.edu/vizier.php/.
[25] NED Batch Jobs, http://nedwww.ipac.caltech.edu/help/batch.html/.
[26] CasJobs, http://casjobs.sdss.org/CasJobs/.
[27] TOPCAT, http://www.star.bris.ac.uk/~mbt/topcat/.
[28] 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.

[29] Clive Page, Indexing the Sky, http://wiki.astrogrid.org/bin/view/Astrogrid/SkyIndexing/.
[30] Hierarchical Triangular Mesh, http://taltos.pha.jhu.edu/htm/.
[31] Hierarchical Equal Area isoLatitude Pixelisation, http://www.eso.org/science/healpix/.
[32] 刘超, 基于虚拟天文台的数据挖掘技术及其在银河系晕结构研究中的应用. [博士学位论文], 北京: 中国科学院国家天文台, 2008.
[33] Globus Alliance, http://www.globus.org/.
[34] Weir, N., Fayyad, U., et al., The SKICAT System for Processing and Analyzing Digital Imaging Sky Surveys, Publications of the Astronomical Society of the Pacific, 1995, 107, 1243.
[35] Weir, N., Fayyad, U., Djorgovski, S. G., Automated Star/Galaxy Classification for Digitized Poss-II, Astronomical Journal, 1995, 109, 2401.
[36] Goebel, J., Stutz, J., Volk, K., Walker, H., et al., A Bayesian classification of the IRAS LRS Atlas, Astronomy & Astrophysics, 1989, 222, 5 - 8.
[37] De Carvalho, R., Djorgovski, S. G., et al., Clustering Analysis Algorithms and Their Applications to Digital POSS-II Catalogs, ASP Conference Series, 1995, 77, 272.
[38] Yoo, J., Gray, A., Roden, J., et al., Analysis of Digital POSS-II Catalogs Using Hierarchical Unsupervised Learning Algorithms, ASP Conference Series, 1996, 101, 41.
[39] Connolly, A. J., Szalay, A. S., A Robust Classification of Galaxy Spectra: Dealing with Noisy and Incomplete Data, The Astronomical Journal, 1999, 117, 2052 - 2062.
[40] Hatziminaoglou, E., Mathez, G., Pello, R., Quasar candidate multicolor selection technique: a different approach, Astronomy & Astrophysics, 2000, 359, 9 -17.
[41] Wolf, C., Dye, S., Kleinheinrich, M., Meisenheimer, K., Rix, H. W., Wisotzki, L., Deep BVR photometry of the Chandra Deep Field South from the COMBO-17 survey, Astronomy & Astrophysics, 2001, 377, 442 - 449.
[42] McGlynn, T. A., Suchkov, A. A., Winter, E. L., Hanisch, R. J., White, R. L., Ochsenbein, F., Derriere, S., Voges, W. et al., Automated Classification of ROSAT Sources Using Heterogeneous Multiwavelength Source Catalogs, The Astronomical Journal, 2004, 616, 1284 - 1300.
[43] Carballo, R., Cofino, A. S., Gonzalez-Serrano, J. I., Selection of quasar candidates from combined radio and optical surveys using neural networks, Monthly Notice of the Royal Astronomical Society, 2004, 353, 211 - 220.
[44] Suchkov, A. A., Hanisch, R. J., Margon, Bruce., A Census of Object Types and Redshift Estimates in the SDSS Photometric Catalog from A Trained Decision Tree Classifier, The Astronomical Journal, 2005, 130, 2439 - 2452.

[45] Ball, N. M., Brunner, R. J., Myers, A. D., Tcheng, D., Robust Machine Learning Applied to Astronomical Data Sets. I. Star-Galaxy Classification of the Sloan Digital Sky Survey DR3 Using Decision Trees, The Astronomical Journal, 2006, 650, 497 - 509.

[46] Storrie-Lombardi, M. C., Lahav, O., Sodre, L. Jr., et al. Morphological Classification of Galaxies by Artificial Neural Networks, Monthly Notice of the Royal Astronomical Society, 1992, 259, 8.

[47] Adams, A., Woolley, A., Hubble classification of galaxies using neural networks, Vistas in Astronomy, 1994, 38, 273-280.

[48] Zhang, Y., Zhao, Y., Learning Vector Quantization for Classifying Astronomical Objects, Chinese Journal of Astronomy & Astrophysics, 2003, 3, 183 - 190.

[49] Zhang, Y., Zhao, Y., NB, ADTree and MLP Comparison in Separating Quasars from Large Survey Catalogues, Chinese Journal of Astronomy & Astrophysics, 2007, 7, 289.

[50] Zhao, Y., Zhang, Y., Comparison of decision tree methods for finding active objects, Advances in Space Research, 2008, 41, 1955-1959.

[51] Baum, W., A Photoelectric determinations of redshifts beyond 0.2 c. Astronomical Journal. 1957, 62, 6 - 7.

[52] Baum, W., A Photoelectric Magnitudes and Red-Shifts. Proceedings from IAU Symposium no. 15, Macmillan Press, New York, 1962, p.390.

[53] Fontana, A., DOdorico, S., Poli, F., et al., Photometric Redshifts and Selection of High-Redshift Galaxies in the NTT and Hubble Deep Fields, The Astronomical Journal, 2000, 120, 2206 – 2219.

[54] Connolly, A. J., et al., Slicing Through Multicolor Space: Galaxy Redshifts from Broadband Photometry, Astronomical Journal, 1995, 110, 2655.

[55] Sowards-Emmerd, D., Smith, J. A., McKay, T. A., et al., A Catalog of Photometry for Las Campanas Redshift Survey Galaxies on the Sloan Digital Sky Survey System, Astronomical Journal, 2000, 119, 2598 - 2604.

[56] Wadadekar, Y. Estimating Photometric Redshifts Using Support Vector Machines, The Publications of the Astronomical Society of the Pacific, 2005, 117, 79 – 85.

[57] Collister, A. A., Lahav, O. ANNz: Estimating Photometric Redshifts Using Artificial Neural Networks. The Publications of the Astronomical Society of the Pacific. 2004, 114, 345 – 351.

[58] Firth, A. E., Lahav, O., Somerville, R. S., Estimating photometric redshifts with artificial neural networks. Monthly Notice of the Royal Astronomical Society. 2003, 339, 1195 – 1202.

[59] Vanzella, E., et al., Photometric redshifts with the Multilayer Perceptron Neural Network: Application to the HDF-S and SDSS. Astronomy and Astrophysics. 2004, 423, 761 – 776.

[60] Bentley, J. L., Multidimensional binary search trees used for associative searching. 1975, Commun. ACM 18, 9 (Sep. 1975), 509 – 517.

[61] Kunszt, P. Z., Szalay, A. S., Csabai, I., Thakar, A. R., The Indexing of the SDSS Science Archive, Proceeding of ADASS IX, 2000, 216, 141.

[62] Maneewongvatana, S., Mount, D. M., Analysis of Approximate Nearest Neighbor Searching with Clustered Point Sets, DIMACS Series in Discrete Mathematics and Theoretical Computer Science, 2002, 59, 105.

[63] Hsieh B.C., Yee H.K.C., Lin H., Gladders M.D., A Photometric Redshift Galaxy Catalog from the Red- Sequence Cluster Surve, The Astrophysical Journal Supplement Series, 2005, 158, 161 - 177.

[64] Kubica, J., Denneau, L., Grav, T., Heasley, J., Jedicke, R., Masiero, J., Milani, A., Moore, A,. et al., Efficient intra- and inter-night linking of asteroid detections using kd-trees, Icarus, 2007, 189(1), 151.

[65] Vapnik, V. N., The Nature of Statistical Learning Theory. New York: Springer, 1995.

[66] Wozniak, P. R., Akerlof, C., Amrose, S., Brumby, S., Casperson, D., Gisler, G., Kehoe, R., Lee, B. et al., Classification of ROTSE Variable Stars using Machine Learning, American Astronomical Society, 2001, 33, 1495.

[67] Williams, S. J., Wozniak, P. R., Vestrand, W. T., Gupta, V., Identifying Red Variables in the Northern Sky Variability Survey, The Astronomical Journal, 2004, 128, 2965 – 2976.

[68] Humphreys, R. M., Karypis, G., Hasan, M., Kriessler, J., Odewahn, S. C., Experiments in Automating the Morphological Classification of Galaxies, American Astronomical Society, 2001, 33, 1322.

[69] Qu, M., Shih, F. Y., Jing, J., Wang, H. M., Automatic Solar Flare Detection Using MLP, RBF, and SVM, Solar Physics, 2003, 217(1), 157 - 172.

[70] Zhang, Y., Zhao, Y., Classification in Multidimensional Parameter Space: Methods and Examples, Publication of the Astronomical Society of the Pacific, 2003, 115, 1006 - 1018.

[71] Zhang, Y., Zhao, Y., Automated clustering algorithms for classification of astronomical objects, Astronomy & Astrophysics, 2004, 422, 1113 – 1121.

[72] Wang, D., Zhang, Y. X., Liu, C., Zhao, Y. H., Two Novel Approaches for Photometric Redshift Estimation based on SDSS and 2MASS, Chinese Journal of Astronomy & Astrophysics, 2008, 8, 119 – 126..

[73] Wang, D., Zhang, Y. X., Liu, C., Zhao, Y. H., Kernel regression for determining photometric redshifts from Sloan broad-band photometry, Monthly Notices of the Royal Astronomical Society, 2007, 382, 1601 - 1606.

[74] Rohde, D. J., Drinkwater M. J., Gallagher M. R., Downs, T.,Doyle, M. T., Applying machine learning to catalogue matching in astrophysics, Monthly Notice of the Royal Astronomical Society, 2005, 360, 69 - 75.

[75] Rohde, D. J., Gallagher, M. R., Drinkwater, M. J., Pimbblet, K. A., Matching of catalogues by probabilistic pattern classification, Monthly Notice of the Royal Astronomical Society, 2006, 369, 2 - 14.

[76] Breiman, L., Random Forests, Machine Learning, 2001, 45, 5 - 32.

[77] Breiman, L., Last, M., Rice, J., In: Statistical challenges in astronomy, Third Statistical Challenges in Modern Astronomy (SCMA III) Conference, University Park, PA, USA, July 18-21 2001, Eric D. Feigelson, G. Jogesh Babu (eds.). New York: Springer, 2003, p.243-254.

[78] Albert, J., et al., Implementation of the Random Forest Method for the Imaging Atmospheric Cherenkov Telescope MAGIC, 2007, preprint(astro-ph/0709.3719).

[79] Carliles, S., Budavari, T., Heinis, S., et al., Photometric Redshift Estimation on SDSS Data Using Random Forests, 2007, preprint(astro-ph/0711.2477).

[80] Bailey, S., Aragon, C., Romano, R., et al., How to Find More Supernovae with Less Work: Object Classification Techniques for Difference Imaging, The Astrophysical Journal, 2007, 665, 1246 - 1253.

[81] SDSS, http://www.sdss.org/

[82] York, D. G., et al., The Sloan Digital Sky Survey: Technical Summary, Astronomical Journal, 2000, 120, 1579 -1587

[83] 2MASS, http://pegasus.phast.umass.edu/

[84] Becker, R. H., Helfand, D.J., White, R.L., et al., A Catalog of 1.4 GHz Radio Sources from the FIRST Survey, Astrophysical Journal, 1997, 475, 479.

[85] Voges, W., Aschenbach B., Boller T., et al., The ROSAT all-sky survey bright source catalogue, Astronomy & Astrophysics, 1999, 349, 389 - 405.

[86] Monet, D. G., Levine S. E., Casian B., et al., The USNO-B Catalog, The Astronomical Journal, 2003, 125, 984 - 993.

[87] Vanzella, E., Cristiani, S., Fontana, A., et al., Photometric redshifts with the Multilayer Perceptron Neural Network: Application to the HDF-S and SDSS, Astronomy & Astrophysics, 2004, 423, 761 - 776.

[88] L, Li., Y, Zhang., Y, Zhao., D, Yang., Multi-parameter estimating photometric redshifts with artificial neural networks, Chinese Journal of Astronomy and Astrophysics, 2007, 7(3), 448 - 456.

[89] 王丹, 基于大型巡天数据的测光红移算法研究—算法研究与工具开发. [博士学位论文], 北京: 中国科学院国家天文台, 2007.

[90] Vi, http://www.vim.org/.

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

[92] Warren, S. J., Hewett, P. C., & Osmer, P. S., A wide-field multicolor survey for high-redshift quasars, Z greater than or equal to 2.2. 3: The luminosity function, Astrophysical Journal, 1994, 421, 412 - 433.

[93] Gregg, M. D., Becker, R. H., White, R. L., et al., The First Bright QSO Survey, Astronomical Journal, 1996, 112, 407.

[94] White, R. L., et al., The FIRST Bright Quasar Survey. II. 60 Nights and 1200 Spectra Later, The Astrophysical Journal Supplement Series, 2000, 126, 133 - 207.

[95] Brunzendorf, J., Meusinger, H., A QSO survey via optical variability and zero proper motion in the M 92 field. I. QSO candidates and selection effects, Astronomy & Astrophysics, 2001, 373, 38 - 55.

[96] Meusinger, H., Brunzendorf, J., A QSO survey via optical variability and zero proper motion in the M 92 field. II. Follow-up spectroscopy and properties of the QSO sample, Astronomy & Astrophysics, 2001, 374, 878 - 894.

[97] Wei, J. Y., Xu, D. W., Dong, X. Y., et al., An AGN sample with high X-ray-to-optical flux ratio from RASS. I. The optical identification, Astronomy and Astrophysics Supplement, 1999, 139, 575 - 599.

[98] He, X. T., Wu, J. H., Yuan, Q. R., et al., The Multiwavelength Quasar Survey. I. Initial Results, Astronomical Journal, 2001, 121, 1863 - 1871.

[99] Cheng, Y., He, X. T., Wu, J. H., et al., The Multiwavelength Quasar Survey. II. Quasars in the Coma Cluster, Astronomical Journal, 2002, 123, 578 - 582.

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

[101] Zhang, Y., Zhao, Y., Cui, C., Data Mining and Knowledge Discovery in Database of Astronomy, Progress in Astronomy, 2002, 20(4): 312 - 323.

[102] Zhang, Y., Zhao, Y., Gao, D., Decision table for separating quasars from stars, Advances in Space Research, 2008, 41, 1949-1954.

[103] Zheng, H., Zhang, Y., Feature selection for high dimensional data in astronomy, Advances in Space Research, 2008, 41, 1960-1964.

发表文章 I

（英文期刊文章）

- **Gao, D.**, Zhang, Y., Zhao, Y., Support Vector Machines and Kd-tree for Separating Quasars from Large Survey Data Bases, MNRAS, 2008, 386, 1417
- **Gao, D.**, Zhang, Y., Zhao, Y., Random Forest algorithm for Classifying Multi-Wavelength Data, ChJAA, 2008, submitted (in the second review)
- **Gao, D.**, Zhang, Y., Zhao, Y., A System Integrated with Query, Cross-matching and Visualization， Advanced Software and Control for Astronomy. Edited by Lewis, Hilton; Bridger Alan. Proceeding of the SPIE, 2006, Vol 6274, pp. 627414-1
- **Gao, D.**, Zhang, Y., Zhao, Y., The Application of kd-tree in Astronomy, Proceeding of ADASS XVII, 2008, in press
- Zhang, Y., Zhao, Y., **Gao, D.**, Discrimination of Point Sources Using Radial Basis Function Network and Random Forest, Proceeding of ADASS XVII, 2008, in press
- Zhang, Y., Zhao, Y., **Gao, D.**, Decision table for classifying point sources based on FIRST and 2MASS databases, Advances in Space Research, 2008, 41, 1949-1954
- Liu, C., Wang, D., Liu, B., **Gao, D.**, Cui, C. and Zhao, Y., An astronomical data mining application framework for virtual observatory, Advanced Software and Control for Astronomy. Edited by Lewis, Hilton; Bridger Alan. Proceeding of the SPIE, 2006, Vol 6274, pp. 627415-1

发表文章 II

（中文期刊文章）

- **高丹**, 张彦霞, 赵永恒, 海量多波段型表数据的交叉证认的实现, 天文研究与技术, 2005, 2(3), 186
- **高丹**, 张彦霞, 赵永恒, 天文星表入库的自动化, 天文研究与技术, 2007, 4(1), 53
- **高丹**, 张彦霞, 赵永恒, 中国虚拟天文台交叉证认工具的开发和应用, 天文学报, 2008, in press
- **高丹**, 路勇, 张彦霞, 赵永恒, 海量星表融合系统（XMaS_VO）的设计与开发, 天文研究与技术, 2008, 5(2)，138
- 刘超, 田海俊, **高丹**, 杨阳, 路勇, 崔辰州, 赵永恒, 异地异构天文数据资源的统一访问, 天文研究与技术, 2008, 5(2), 146

致 谢

五年的研究生阶段即将结束，我也要告别长达 19 年的求学生涯，即将步入工作岗位。这五年是我人生最宝贵的时光，我从一个活蹦乱跳刚满 20 岁的孩子成长为一个做好准备进入社会工作的社会人。硕博期间我幸运地获得了中科院朱李月华奖学金，还被评为优秀学生干部和优秀毕业生。这一切的成绩都离不开老师们的指导和朋友们的帮助，我在这里由衷地感谢所有曾经给予我以引导、鼓励、帮助和看着我成长的人。

感谢我的恩师赵永恒研究员。赵老师学识渊博、谦虚诚恳、宽以待人的品格深深地影响着我。我第一次见到赵老师的时候，就由衷地敬佩他平易近人的态度和广博的知识。感谢赵老师这些年来在科研上对我的引导和指点，他对学术前沿有着敏锐的洞察力，每次讨论都让我有豁然开朗的感觉。感谢赵老师的言传身教，让我在为人处事的道理和科研工作的态度等方面收获颇多。我这五年来，无数次看到赵老师忙碌的身影，他为 LAMOST 项目倾注了太多的心血。每次看到原本年轻的赵老师不知不觉中添上的丝丝白发，都让我隐隐心痛。赵老师，您辛苦了！

感谢我的课题指导老师张彦霞副研究员。张老师在学术上有着扎实的基础，治学勤奋、踏实、严谨，对科研工作有着执着和钻研的精神，对课题有丰富的想法和见解。让我印象最深的就是她办公室里的一堆堆文献资料和她伏案学习的身影。张老师为人非常热情、坦诚，无私地帮助别人，她也总是能无私地为学生的利益考虑，站在学生的角度去理解。我是张老师指导的第一个学生，也是她指导时间最长的学生。感谢张老师这些年来对我的严格要求，督促我学习进步，锻炼了我科研的思维能力，培养了我踏实的工作态度。我的每一点成绩都离不开张老师的心血，我的每一篇文章、每一次做的报告都离不开张老师耐心的指导，她总能让我收获和提高。感谢张老师在生活上、思想上和为人处事方面给予我的数不

清的照顾、引导和提醒。感谢张老师给予还处于不太懂事、天真的年纪的我的包容和理解。张老师对我付出得太多太多，不是几句感谢所能表达的。

感谢国家天文台赵刚研究员，感谢赵老师在我刚入学时对我研究方向的指引和在我求学过程中对我的帮助。

感谢虚拟天文台的崔辰州博士对我的关心和指导，感谢他每一次辛苦组织的国际和国内会议，总让我受益匪浅。

感谢张洪起研究员在我刚来天文台时对我的热情指导，感谢胡景耀研究员、邓李才研究员、魏建彦研究员、武向平研究员、韩金林研究员、南仁东研究员、陈学雷研究员，他们严谨的治学态度和人格魅力深深感染了我。

感谢国家天文台的彭勃老师、姜晓军老师、徐达维老师、林钢华老师 、梁艳春老师、陆烨老师、董素珍老师、马艳霞老师、朱爱萍老师、赵景芝老师的热心帮助。

感谢北京师范大学的何香涛教授、北京大学的吴学兵教授、中科院自动化所的吴福朝研究员、国家天文台的邓李才研究员、周旭研究员、吴宏研究员、马骏研究员，感谢他们百忙之中参加我的论文答辩，并对我的论文给予的建议和指导。

感谢研究生办公室的杜红荣老师和艾华老师，感谢她们热情而细致的工作，感谢她们对我多方面的照顾和理解。

感谢党支部的孙盛慈老师、李金增老师、张彦霞老师和罗阿理老师，每一次和你们的谈话都让我收获很多，感谢孙老师在我最脆弱的时候给我的温暖和帮助，感谢王宜副台长在我入党前对我的谈话指导。

感谢刘晓群书记、孙超老师、姜晓军老师、王晓倩、张晓宇、徐刚和翟萌对我的帮助。

感谢 LAMOST 的老师：陈英老师、袁晖老师、石火明老师、陈建军老师、张昊彤老师、王钢老师、门力老师、曹淑蕴老师、李颀老师、冯磊老师的帮助。

感谢 LAMOST 的刘超师兄和中科院声学所的张海滨，他们在软件开发方面给了我很大的指导和帮助，让我对软件工程有了更深入的理解。

感谢和我一起探讨过课题并给予我帮助的杨雁宾博士、曹晨和上海天文台齐

朝祥。

感谢英国 Astrogrid 的秦岭女士和美国加州理工学院的 Lijun Zhang 女士在伦敦对我的照顾和帮助，让第一次一个人身处异国他乡的我倍感亲切和温暖。

感谢 LAMOST 的朋友：李会贤、桑健、刘波、贾磊、吴潮、邹思成、王淑青、张健楠、尹红星、李丽丽、薛元、陈晓艳、吴悦、郑征、孙华平、孙士卫、罗宇、杨阳、田海俊、宋轶晗、王凤飞、杨三军、杨帆、严太生、姚嵩、何勃亮、彭南博、邢丽君、闫岩。

感谢天文台的朋友：李晓昕、朱轶楠、黄静、曹晨、高裕、郭铨、陕欢源、毛羽丰、胡晨、王陈、段帷、屈艳、王汇娟、王二超、徐海清、张丽云、高健健、文中略、张惠、李丹、张蕾、黄浩、刘凤山、陈洁、刘玉娟、夏方、李蓉、张逢春、张灵娣、赵景昆、郑伟康、范舟、杨尚斌、朱兰、肖江。

感谢中科院研究生院博士合唱团，感谢西部行演出，帮助我从人生的低谷中走出来。感谢马石庄院长、刘团长、林玉赤老师的照顾，感谢合唱团的每一次演出，感谢合唱团的所有团员，他们是白云祥、张轶群、蔡世淳、胡毅庆、王敏、陈名、陈果、张晗、王艺瑾、陈兰、冯晓颖、罗卫军、曹福祥、黄芝、周洁、吴开明、孙朝阳、陈寨伟、姜靖、张鹏、杜磊、李卓、张文涛、王嵩、王莹、王建芳、黄彤明。

感谢刘红雨老师的鼓励和支持，让我于四年前创办中科院研究生院心理协会并坚持至今，实现了自己的梦想。感谢刘蓉晖老师、叶健、舒航、刘国卿、屈娥、路勇在协会艰难时对我的帮助。感谢支持心理协会的白京生大夫，教会我不要轻信传言，一句话经过几个人传播会完全偏离本意。感谢心理协会所有的朋友，他们是李宁、黄娆、胡运飞、李卓、李菲、熊金、林皓、王宇、黄杰文、胡继伟、高俊涛、张凌、刘志超、杜贵平、任嘉、覃剑、陈义祥、徐杰、李小青、刘家平。感谢心理学，帮助别人的过程中最大的受益者是我自己。

感谢中科院研究生院的朋友：范丽捷、陈静、刘赵杰、候秋菊、周阳、杜晨光、张玉明、雷晔灿、孙东霞、李涛、郭瑞杰、韩茹。

感谢中科院研究生院职业发展协会、清华心理协会、北师大心理系、中科院

心理所所有帮助过我的老师和朋友们。

感谢我的挚友和所有帮助过我的朋友们。

感谢我的爱人路勇，感谢他的勇敢、付出和包容，感谢他能和我深入沟通并一起成长，感谢他陪伴我获得了学业和爱情的双丰收，下一步我们将为事业、家庭和生活而共同努力。感谢路勇的父母和姐姐一直以来对我的关心和支持。

感谢我的母亲对我无私的爱和支持，感谢她为家庭付出的一般人难以承受的辛苦，感谢她教会了我善良、坦诚、包容和爱。感谢我在天堂的父亲，感谢他在病床前对我遇到挫折的最后的鼓励，却没能亲眼看到我的今天和未来。

高丹

2008-5-8