天津大学 Tianjin University
智能与计算学部 College of Intelligence and Computing

面向智能分析的天文时序数据处理关键技术研究

Research on the key technology of astronomical time series data processing for intelligent analysis

报告人：李琨

2023年4月21日

Contents

天津大学
Tianjin University

目 录

Chapter 1

天津大学
Tianjin University

1 研究背景和意义

Background and Significance

光变曲线智能分析

50 New Planets Confirmed in Machine Learning First – AI Distinguishes Between Real and "Fake" Planets

TOPICS: Algorithm Artificial Intelligence Astronomy Exoplanet Kepler Machine Learning

University Of Warwick

By UNIVERSITY OF WARWICK AUGUST 26, 2020

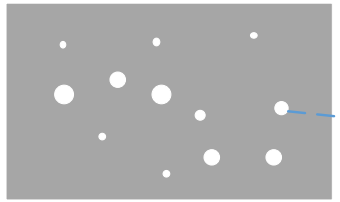
- New machine learning algorithm designed by astronomers and computer scientists from University of Warwick confirms new exoplanets in telescope data
- Sky surveys find thousands of planet candidates, and astronomers have to separate the true planets from fake ones
- Algorithm was trained to distinguish between signs of real planets and false positives
- New technique is faster than previous techniques, can be automated, and improved with further training

光变曲线是源自天文望远镜观测的时间序列数据，记录了天体属性随时间的变化，是系外行星、超新星等时域天文学研究的核心对象。

巡天项目的星表数据量

项目名称	天体条目数
2MASS	470,992,971
WISE	563,921,584
AllWISE	747,634,026
SDSS DR9	1,231,051,050
Gaia DR2	1,692,919,135
PanSTARRS DR1	1,919,106,885

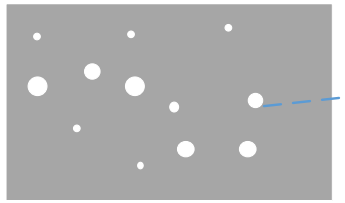
光变曲线构造



天文图像1

ID	RA	Dec	P	…
1	r_1	d_1	p_1	…
2	*r_2*	*d_2*	*p_2*	…
3	r_3	d_3	p_3	…

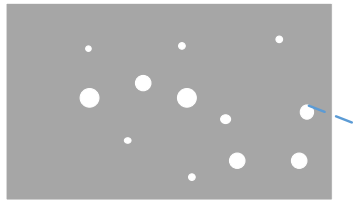
星表1



天文图像2

ID	RA	Dec	P	…
1	*r_1*	*d_1*	*p_1*	…
2	r_2	d_2	p_2	…
3	r_3	d_3	p_3	…

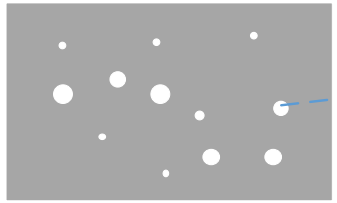
星表2



天文图像3

ID	RA	Dec	P	…
1	r_1	d_1	p_1	…
2	r_2	d_2	p_2	…
3	*r_3*	*d_3*	*p_3*	…

星表3



天文图像4

ID	RA	Dec	P	…
1	*r_1*	*d_1*	*p_1*	…
2	r_2	d_2	p_2	…
3	r_3	d_3	p_3	…

星表4

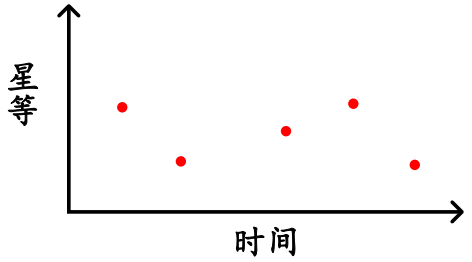
证认计算

$$\sqrt{((r_1 - r_2) \times \cos((d_1 + d_2)/2))^2 + (d_1 - d_2)^2} \le 3\sqrt{R_1^2 + R_2^2}$$

R1 和 R2 是误差半径

光变曲线数据集

TID	RA	Dec	P_T_1	P_T_2	P_T_3	P_T_4	…
1	r_1	d_1	P_1, T_1	P_2, T_2	P_3, T_3	P_4, T_4	…
2	*r_2*	*d_2*	*P_1, T_1*	*P_2, T_2*	*P_3, T_3*	*P_4, T_4*	…



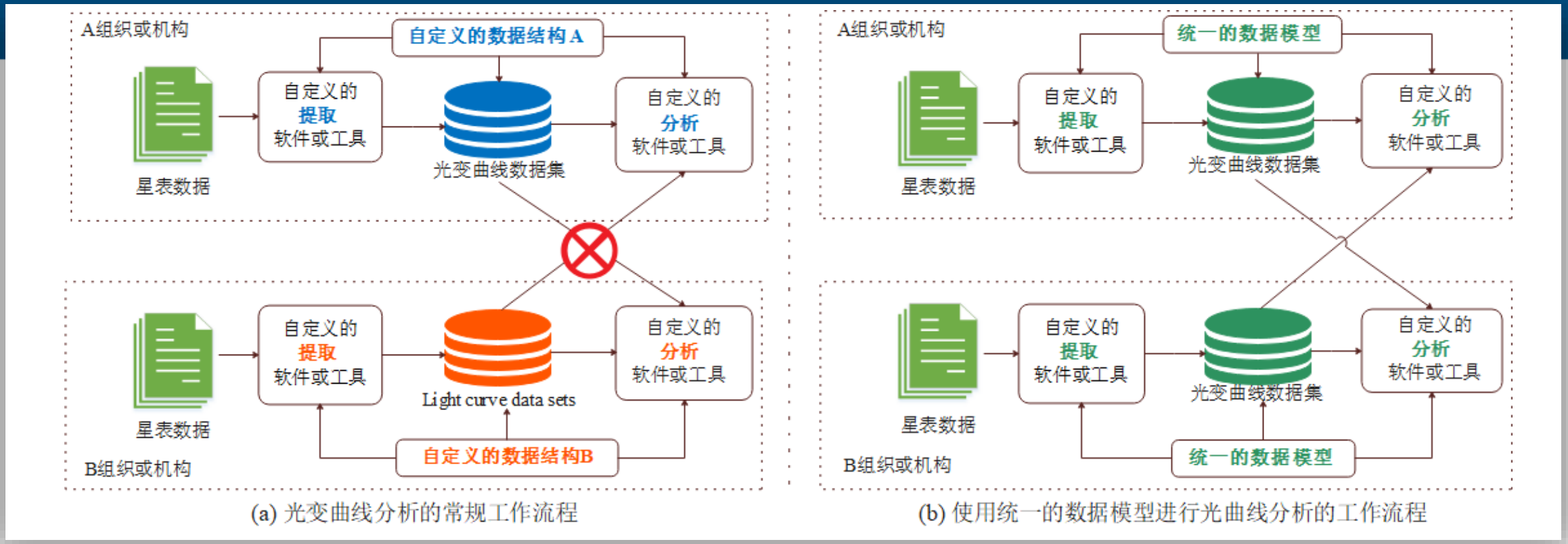
按需检索的方式获得指定天体的光变曲线数据

对天体时序变化的研究仅能限于少量特定候选目标

观测数据的潜在科学价值无法充分挖掘

统一的光变曲线数据模型

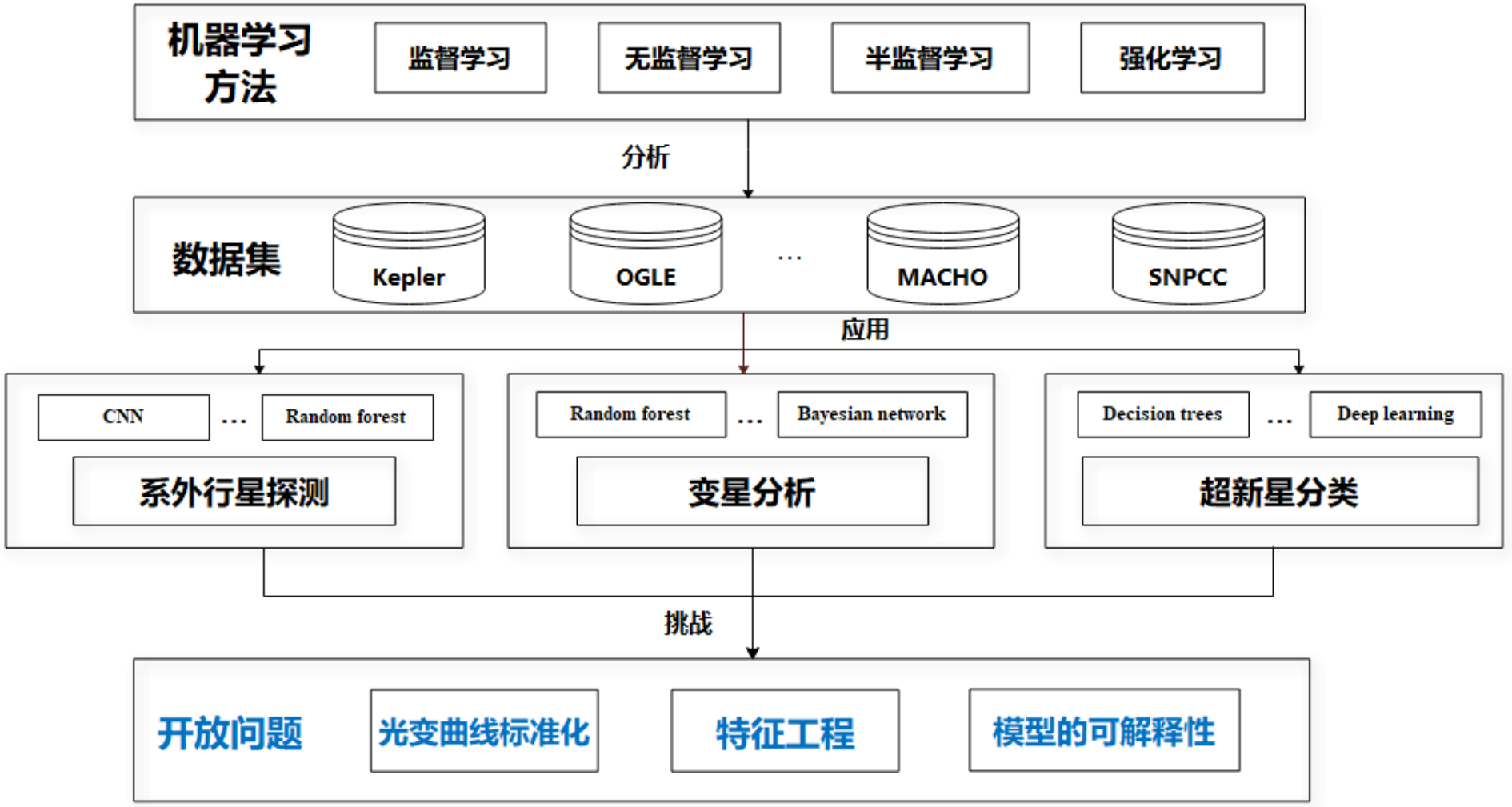
- 不同研究组织发布的光变曲线数据集无法直接应用于别的应用
- 时域天文学团队研究光变曲线分析开发的软件和工具都具有深远的影响



(a) 光变曲线分析的常规工作流程

(b) 使用统一的数据模型进行光曲线分析的工作流程

基于机器学习的光变曲线分析



目前面对的挑战有

1. 光变曲线标准训练集的创建
2. 特征工程
3. 可解释性

Chapter 2

天津大学
Tianjin University

2 研究内容
Research Contents

科学目标

- 建立光变曲线智能分析的数据基础和技术基础并开展示范应用

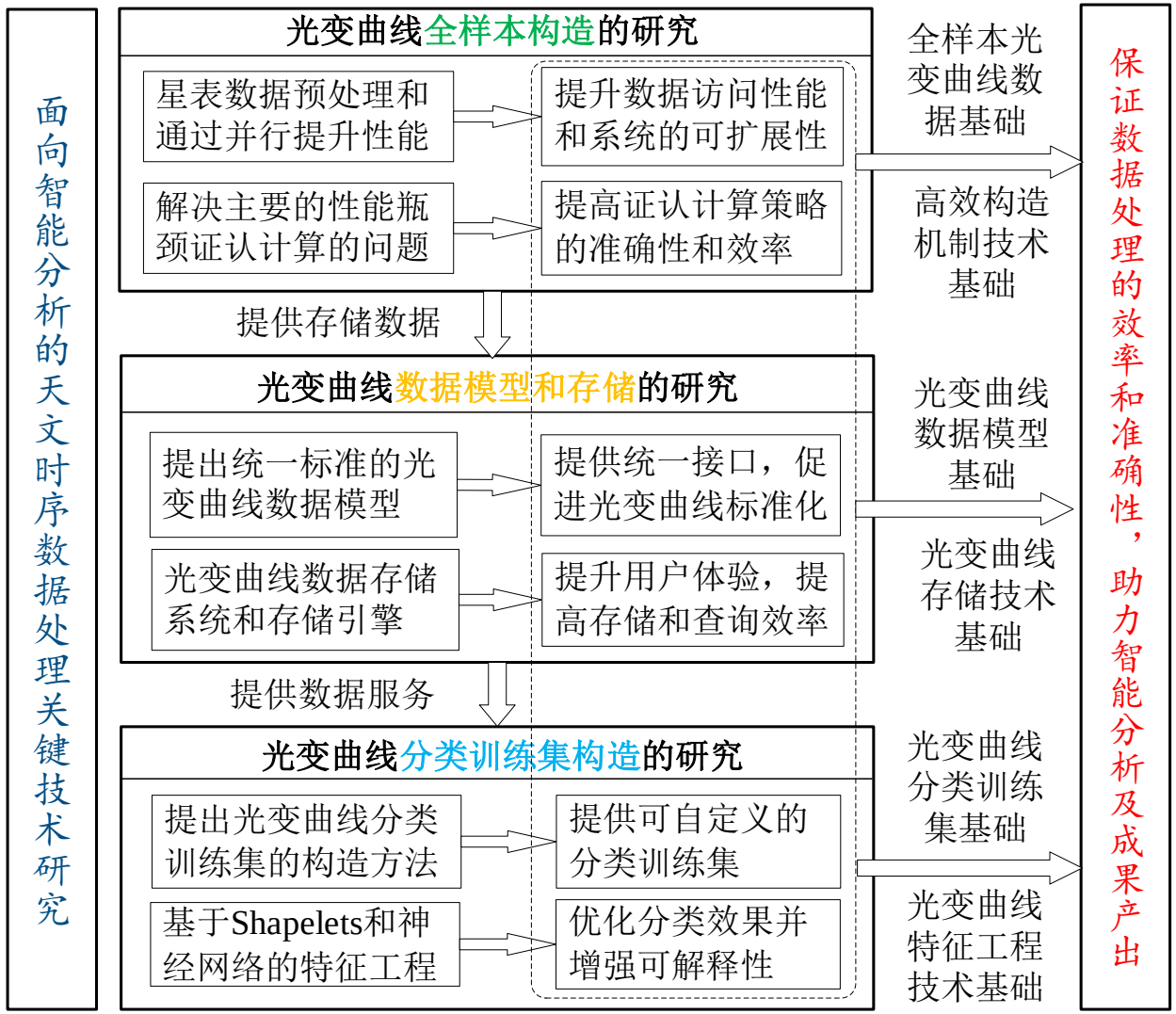
数据基础

- 全样本的光变曲线数据集
- 统一光变曲线数据模型
- 光变曲线分类训练集

技术基础

- 基于证认基准表的星表自证认算法
- 光变曲线自主可控的解耦合存储方案
- 光变曲线特征工程及分类方法

具体研究内容与关键问题



- 光变曲线**构造**研究：

  关键问题：1. 均衡负载的任务划分问题

  2. 星表自证认高效计算问题

- 光变曲线**存储**研究：

  关键问题：1. 光变曲线数据集不兼容问题

  2. 数据采样不均衡的存储问题

- 光变曲线**分析**研究：

  关键问题：1.光变曲线分类的特征提取问题

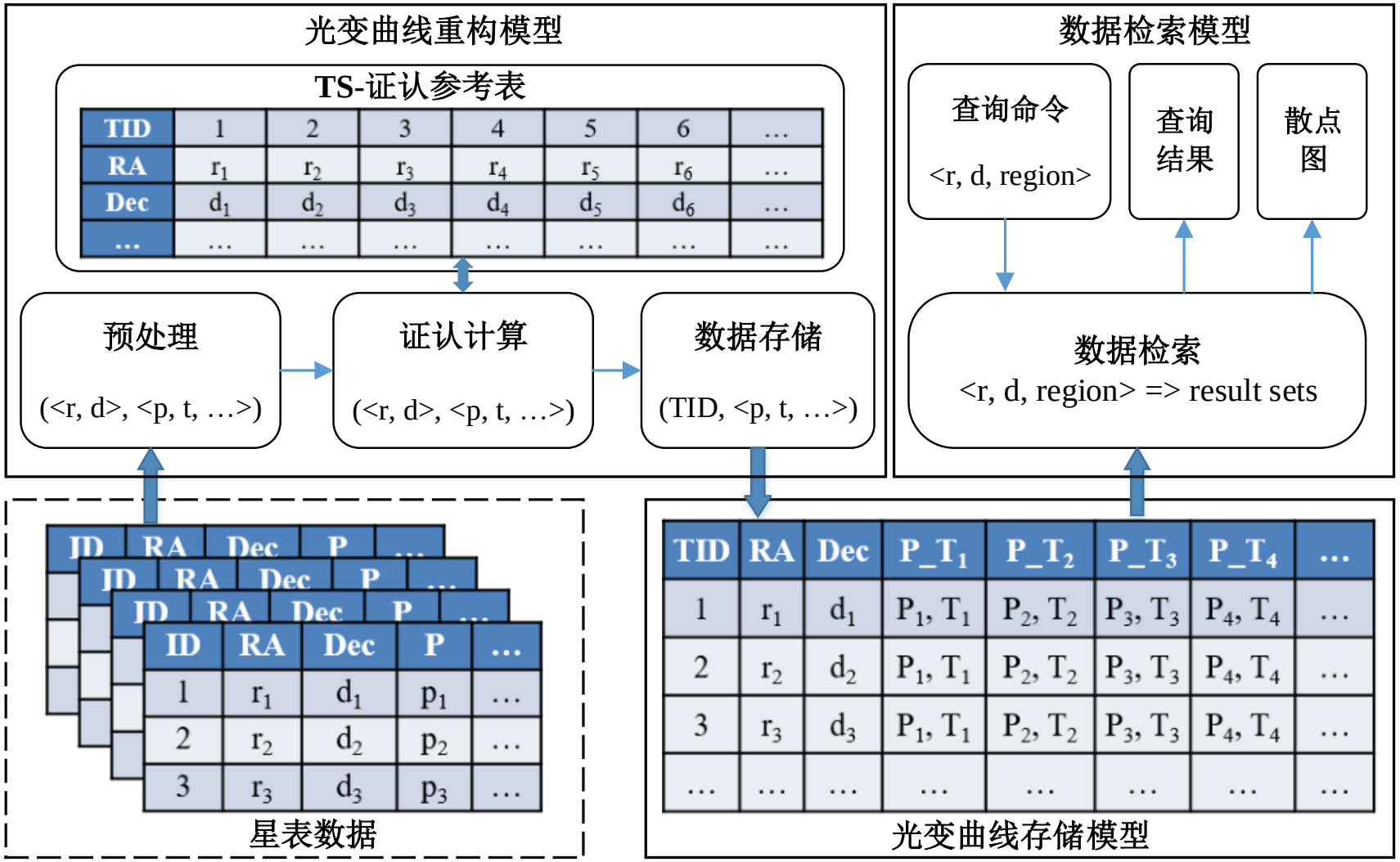
  2.光变曲线分类的效果方面问题

Chapter 4

天津大学
Tianjin University

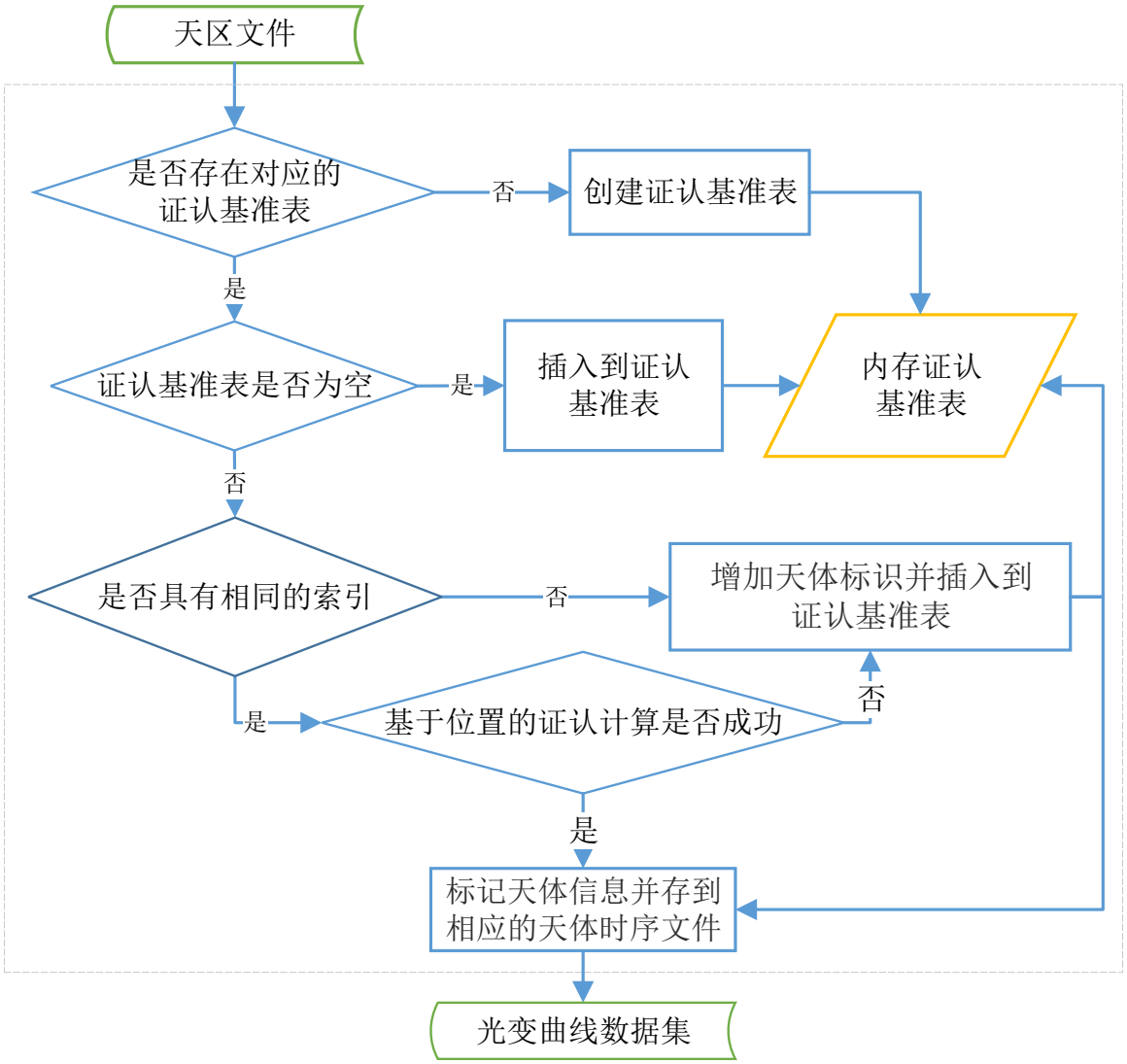
3 研究成果
Research Results

光变曲线全样本数据集的构造



1. 针对星表数据预处理问题，提出了ETL预处理方式。
2. 针对任务分配问题，转化为分步的背包问题，提出了基于动态规划任务划分方法。
3. 针对星表证认计算问题，提出了基于证认基准表的证认方法。

证认算法和任务划分算法



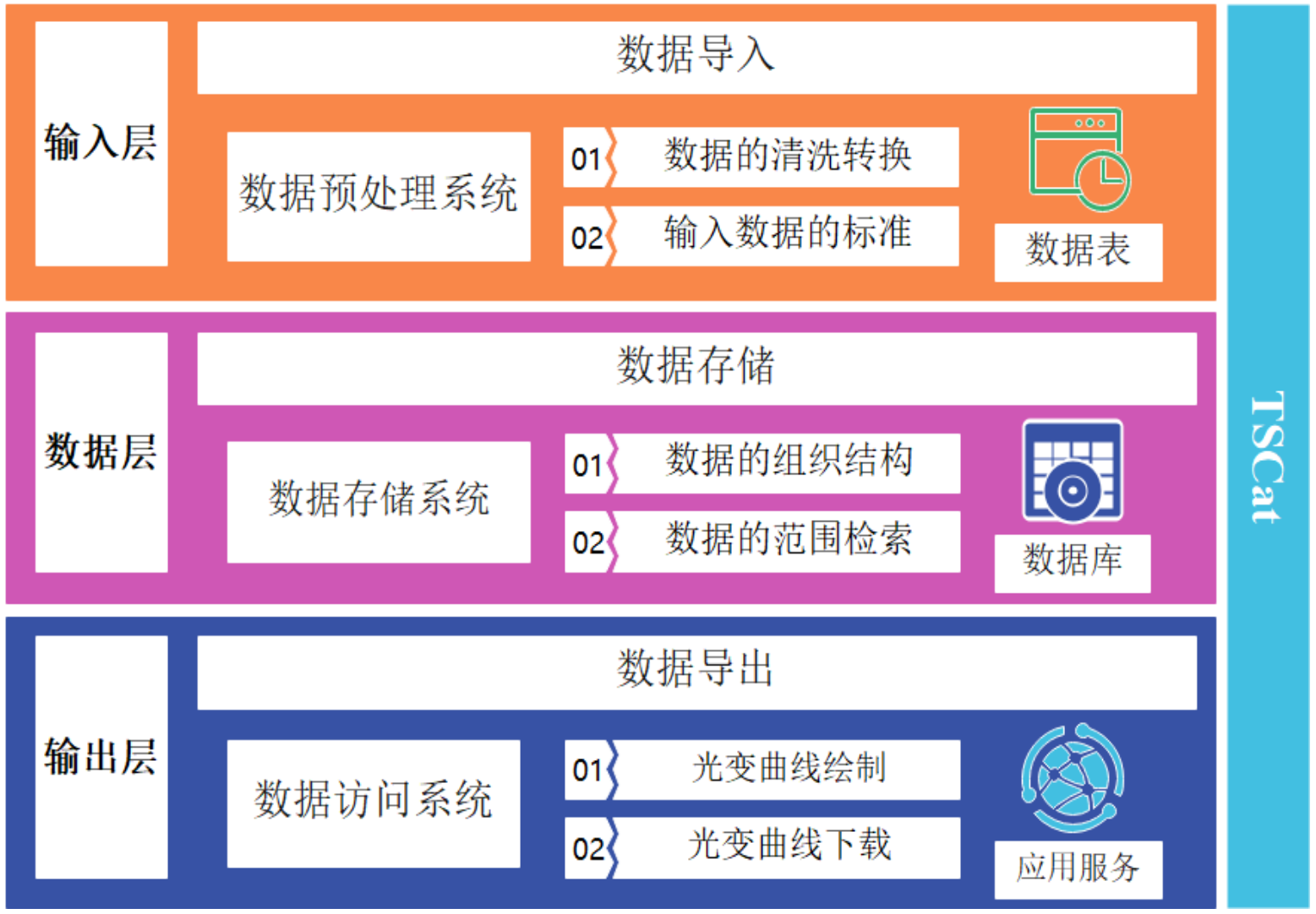
算法 1 任务分配算法

输入: N, M // N 是天区文件编号的集合，M 是给定的进程数

输出: P // 每个进程分配到的天区文件编号的映射

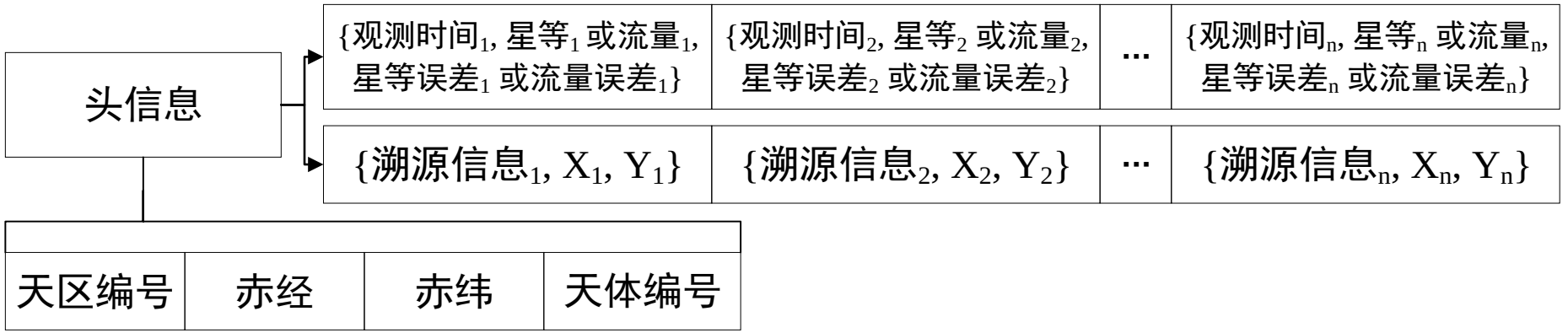
```
P ← {} 将映射关系初始化为空
if 天区文件数 < M then
    P ← 每个进程分配一个天区文件
    return P
end if
f(N) ← 获取每个天区文件的大小
D ← Sort(N, f(N)) 对天区文件进行排序
for 每一个 l ∈ [1, . . . , N] do
    u ← 计算出平均值
    for 每一个 t ∈ D do
        for j = t to u do
            dp[j]=max{dp[j], dp[j-t]+t}
            记录每一个天区文件被分配的进程编号
        end for
    end for
    更新负载均值 u
    删除已分配的天区文件
end for
return P
```

光变曲线模型和存储



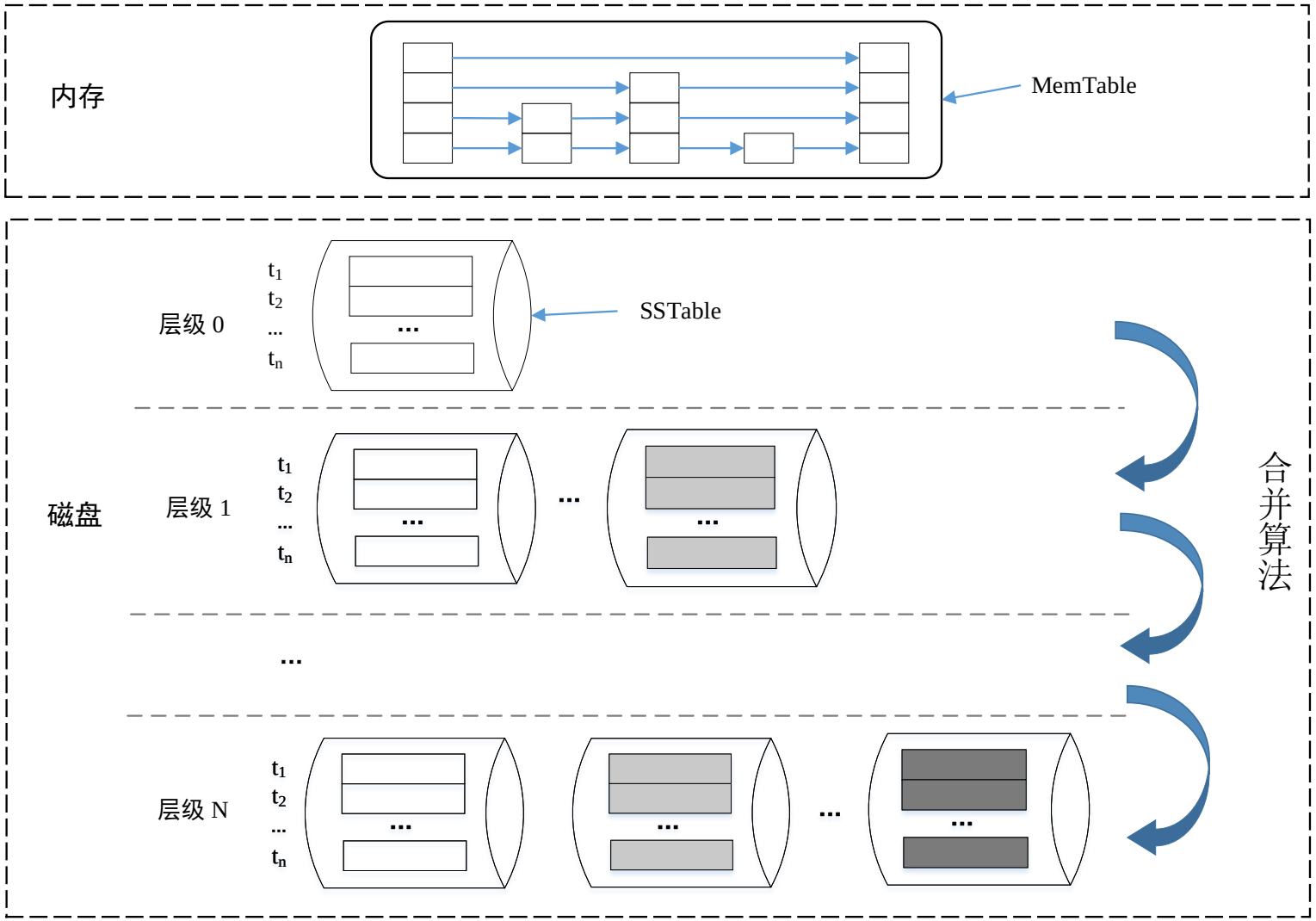
1. 针对光变曲线数据集不兼容问题，提出了TSCat数据模型。
2. 针对光变曲线数据采样不均衡等存储问题，提出了自主可控的解耦合存储方案。并基于TSCat数据模型设计实现了光变曲线存储系统。
3. 针对光变曲线时序元组的读写性能问题，提出了基于LSM树的存储引擎。

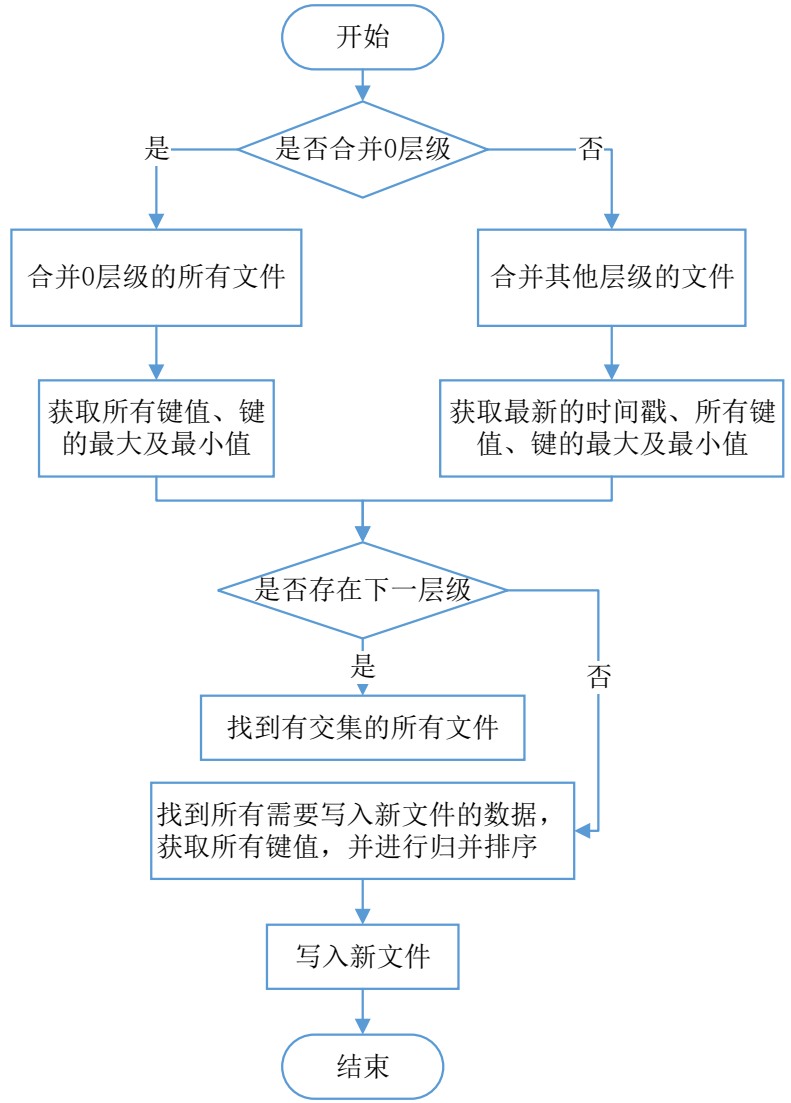
光变曲线模型TSCat



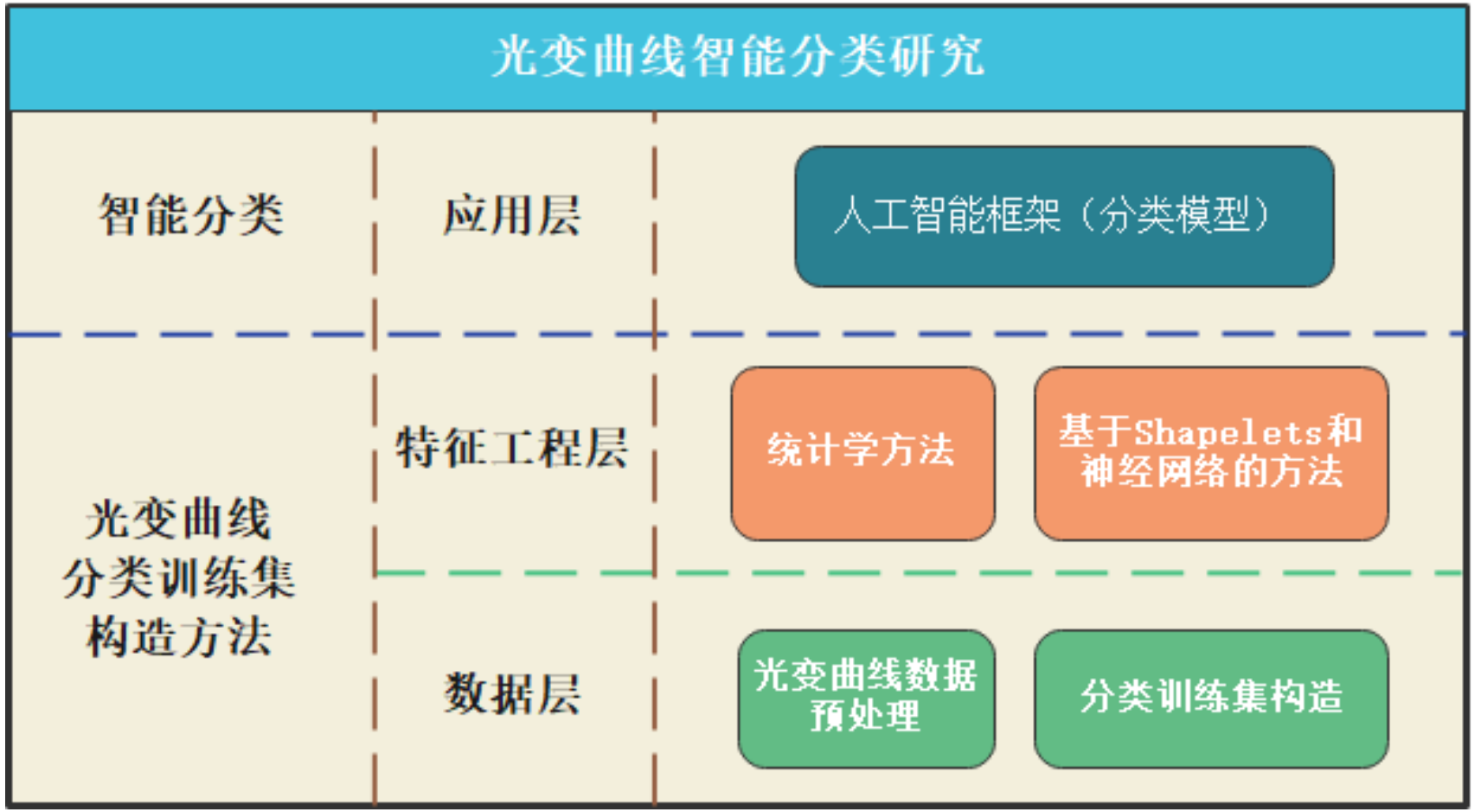
需要的数据信息	光变曲线特征
星等或流量	ampl, fpr, mad, mbrp, pa, pdfp, pst, sk, std, Mean
星等或流量, 时间	Freq{i}_harmonics_amplitude_f{j}, lt, ms, SlottedA_length, Eta_e
星等或流量, 星等误差或流量误差	b1std, StetsonK
星等或流量, 时间, 星等误差或流量误差	CAR_mean, CAR_sigma, CAR_tau

光变曲线存储引擎


内存
MemTable
磁盘
层级 0
层级 1
层级 N
SSTable
合并算法

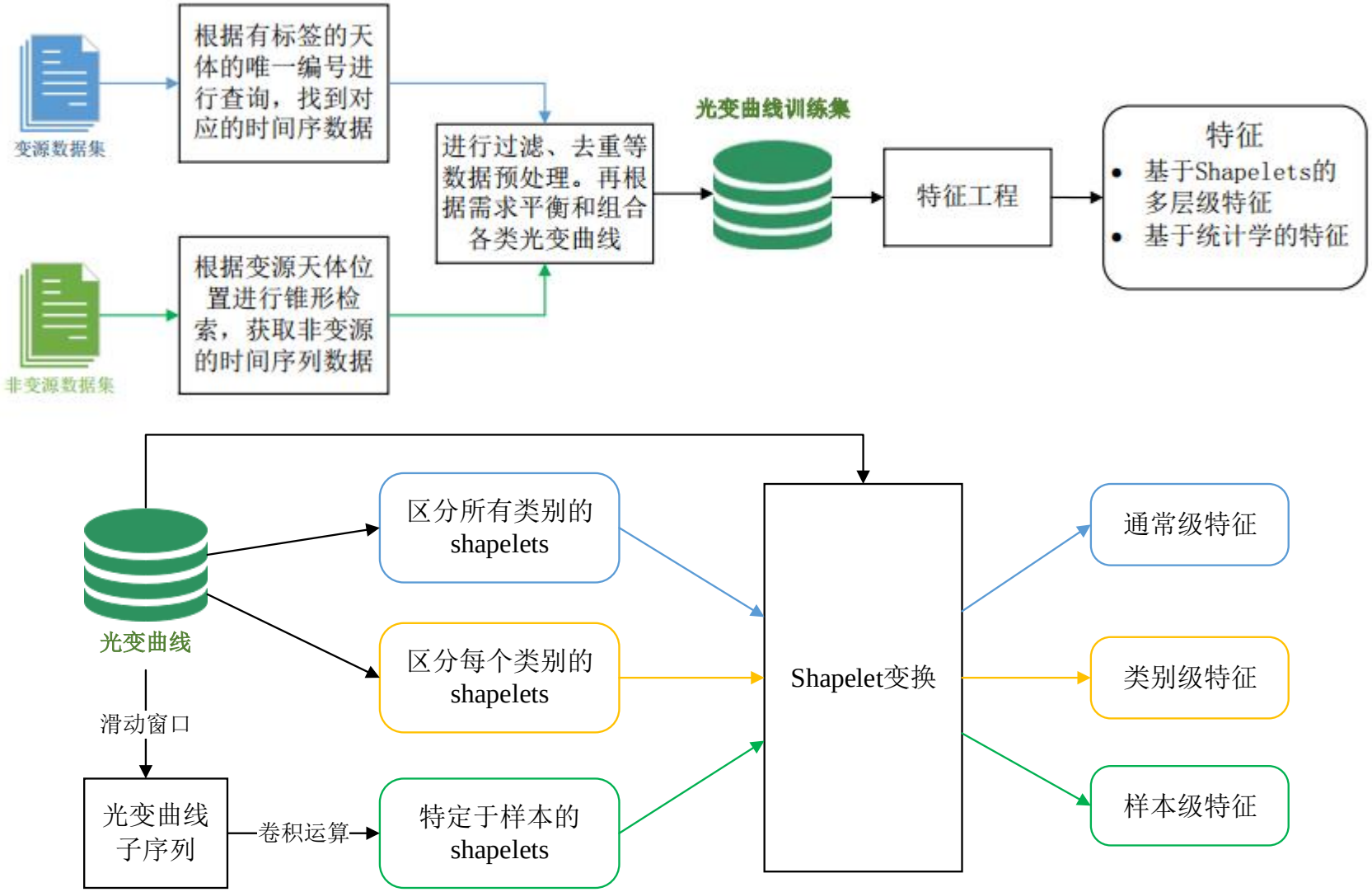

开始
是否合并0层级
是
否
合并0层级的所有文件
合并其他层级的文件
获取所有键值、键的最大及最小值
获取最新的时间戳、所有键值、键的最大及最小值
是否存在下一层级
是
否
找到有交集的所有文件
找到所有需要写入新文件的数据，获取所有键值，并进行归并排序
写入新文件
结束

光变曲线分类训练集的构造



1. 针对训练集短缺的问题，提出了可自定义的光变曲线分类训练集。
2. 针对光变曲线特征提取方面问题，提出了基于shapelets和神经网络的特征工程。
3. 针对分类效果方面的问题，提出了基于多层次的shapelets特征的分类方法。

光变曲线构造方法


变源数据集
根据有标签的天体的唯一编号进行查询，找到对应的时间序数据
非变源数据集
根据变源天体位置进行锥形检索，获取非变源的时间序列数据
进行过滤、去重等数据预处理。再根据需求平衡和组合各类光变曲线
光变曲线训练集
特征工程
特征
基于Shapelets的多层级特征
基于统计学的特征
光变曲线
区分所有类别的shapelets
区分每个类别的shapelets
滑动窗口
光变曲线子序列
卷积运算
特定于样本的shapelets
Shapelet变换
通常级特征
类别级特征
样本级特征

Chapter 6

天津大学
Tianjin University

4 总结与展望
Summary and Outlook

4 总结与展望
Summary and Outlook

光变曲线数据集的构造

提出并实现了一种光变曲线高效构造机制及工具AstroCatR，自海量归档的天文星表数据中，构造出全样本的光变曲线数据集。

光变曲线数据模型与存储

提出了光变曲线数据模型TSCat和自主可控的解耦合存储方案，并设计实现了光变曲线数据存储系统与引擎。

光变曲线分类训练集的构造

提出了光变曲线分类训练集的构造方法和基于Shapelets的特征工程，并基于开源数据构造出了一个可自定义的的光变曲线分类训练集。

4 总结与展望
Summary and Outlook

天津大学
Tianjin University

分布式构造

面向光学时域巡天望远镜阵列的数据处理，提出分布式星表光变曲线数据的构造方案，以期解决多望远镜协同的天文光变曲线数据构造问题。

地理分布式存储

研发实现适于地理分布式光学望远镜阵的分布式存储系统，让各望远镜观测所得的星表数据能够及时自动增量更新至相应天体各自的光变曲线，支持时域天文学领域的重大科学问题研究。

智能分析框架

研发光变曲线智能分析框架，支持多样化的光变曲线智能分析应用和研究，进一步推动人工智能技术方法在光变曲线智能分析中的应用和发展。

天津大學
Tianjin University

智能与计算学部
College of Intelligence and Computing

敬请批评指正！

Thanks for your criticism and correction!