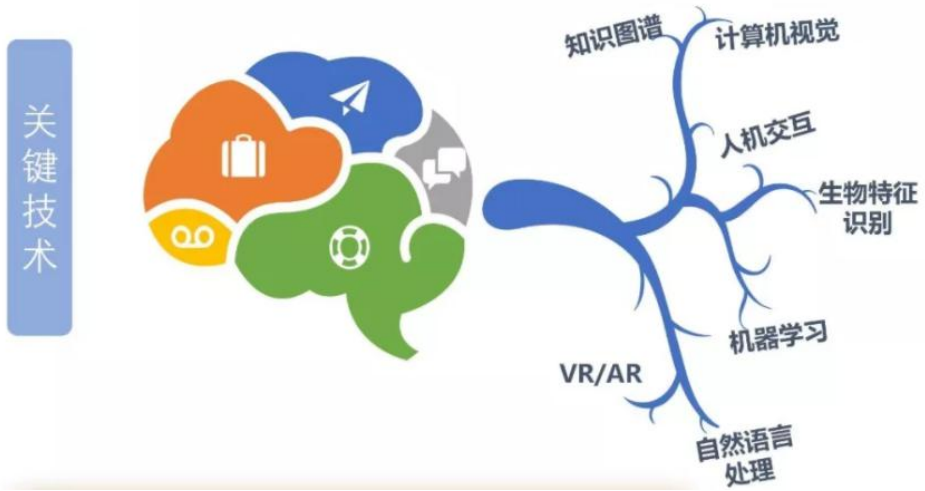
人工智能实验室项目建设方案



目录

一、项目概述

1、项目内容

建设**人工智能实验室项目**，一是利用**人工智能实验室项目**的教学功能提高学生实践能力。二是利用**人工智能实验室项目**的科研功能提高教师在机器学习方向的深度学习研究、人工智能系统开发等能力。三是通过建设**人工智能实验室项目**引进社会技术力量，通过对教师的实践能力培养提升我院学术成果和服务社会的水平。

2、预期效益

通过建设人工智能实验室项目，使学生掌握主流的机器学习算法、计算、分析处理技术，以及深度学习技术，分享成功的人工智能行业应用项目实施经验，学习人工智能应用项目解决方案咨询服务，通过行业背景的资源包提升实用技术，培养创新创业能力。满足工程认证对计算机专业毕业生能力的要求。促进教师机器学习、人工智能方向工程能力的提高。实现科研、教学与社会服务的顺畅衔接，逐步打造以培养人工智能为核心人才的科研、教学基地。

二、项目背景

1）人工智能实验室筹备建设，研究前沿科技未来发展趋势

人工智能是什么？斯坦福大学人工智能研究中心尼尔逊教授提出“人工智能是关于知识的学科——怎样表示知识以及怎样获得知识并使用知识的科学”。美国麻省理工学院的温斯顿教授提出“人工智能就是研究如何使计算机去做过去只有人才能做的智能工作”。人工智能是研究人类智能活动的规律，构造智能的人工系统，研究如何让应用计算机的软硬件来模拟人类某些智能行为的基本理论、方法和技术。人工智能是一门边沿学科，属于自然科学、社会科学、技术科学三向交叉学科，涉及到计算机科学、心理学、医学、数学、哲学和语言学等多个学科。

21世纪各国都高度重视人工智能技术发展，积极布局人工智能领域，而脑科学发展更是人工智能领域的战略制高点。脑科学(brain sciences)是研究人脑的结构与功能的综合性学科，是人工智能的基础。2013 年 4 月 2 日，美国总统奥巴马宣布拨款 1 亿美金，启动脑科学计划，欧盟、日本随即予以响应，分别启动欧洲脑计划以及日本脑计划。由中国科技部、国家自然基金委牵头的“中国脑计划终于在 2015 年 10 月 24 日上线。中国脑计划主要有两个研究方向：1）以探索大脑秘密、攻克大脑疾病为导向的脑科学研究。

2）以建立和发展人工智能技术为导向的类脑研究

国家对人工智能一直很看好，7月份国务院印发《新一代人工智能发展规划》，而最近又有一个重大的政策力挺人工智能，据国家发改委网站消息，为贯彻落实“十三五”规划《纲要》，2018年，国家发展改革委将组织实施“互联网+”、人工智能创新发展和数字经济试点重大工程。

行业分析人士表示，当前人工智能技术已由机器视觉、深度学习等基础研究领域演进到无人驾驶、智能家居等应用领域。随着国家层面对该领域发展的高度重视与支持，各类资本对这一产业的加速布局，未来人工智能发展应用的广度和深度将大大增加。

人工智能的关注度也拉动了人工智能板块的活跃，有人预计 2020年全球AI市场规模将达到1190亿

元，年复合增速约 19.7%；同期，中国人工智能年复合增速将超 50%，远高于全球增速，市场前景广阔。

3）建设人工智能实验室的意义及价值

目前大学现有的实验条件已不能满足以上的实际需求，主要表现在以下几个方面：

1. 实验教学的仪器性能落后

IT 行业是更新换代最快的工程领域之一，大部份 IT 设备的更新换代时间在 3 年以内， 而学院有些机器设备已服役 5、6 年之久，完全不能适应大数据时代对数据存储和运算的要求。

2. 设备种类不能满足要求

人工智能和云计算对数据存储、运算能力要求非常高，常用的计算机实验机房设备不能满足这些需求，而原有实验室是以编程等方向课的需求为基础设计的，因此在服务器、交换机等核心设备上不能满足需求。

3. 缺少相应的支撑软件

深度学习、人工智能、大数据和云计算平台需要有相应的支撑软件环境，而原有的实验室没有这方面的设置。

4. 缺少面向工程能力培养的仿真环境

工程认证要求学生具有工程实践能力，具有解决复杂工程问题的能力这些能力培养实际的仿真平台进行训练，在 2015 版培养方案中已经提出了要求，而现在的实验环境不能支撑这些要求。

综上所述，建设**人工智能实验室项目**是十分必要的。主要体现以下几点：

1. 工程技术人材培养目标的要求

工程认证要求毕业生具有工程实践经验和解决复杂工程问题的能力，特别需要能够提供仿真或基于行业应用案例的实验环境，从以往的实验室建设情况来看，突出的是验证性、演示性实验，而对于具有工程背景的实验环境认识不够、投入不足，导致这方面一直存在短板。通过本项目的实施，可以有效地解决这方面的问题。

2. 创新能力培养的要求

以“大众创业、万众创新”为背景的大学生创业计划，以及大学生创新创业大赛等突出了对创新能力的要求。但在人工智能应用的环境下，如果没有合适的平台和环境，甚至无法真正的进行学习和开发，创新也就无从谈起。本项目的实施解决了这一问题，从物质上保证了基于互联网+的大学生创新研究。

三、项目建设目标

人工智能实验室项目建设目标是：购置支持深度学习、人工智能研究方向实验实训的软硬件设施和实验实践训练数据包，建立支撑学院师生开展人工智能教学与创新研究的实践基地，使计算机学院现有实验室设备上一个新台阶。通过购置国内外先进的实验设备和软件， 提高实验教学和科学研究的水平，充分保证教学和创新训练的需要。

通过本项目的建设，满足我院深度学习及人工智能方向及相关领域的教学实践要求，满足工程能力培养中对毕业生能力，特别求解复杂工程问题能力的要求。教学计划中实验开出率达到100%；新增综合性实验，且综合性实验占总实验学时数不低于50%；新增实训环节100学时以上；提供不少于 3个人工智能项目案例。增强学

生的创新能力，为大学生创新创业计划提供实践平台；为互联网+大赛、大学生程序设计大赛等实践类比赛项目提供训练环境，满足至少 100 人同时在线的性能要求。

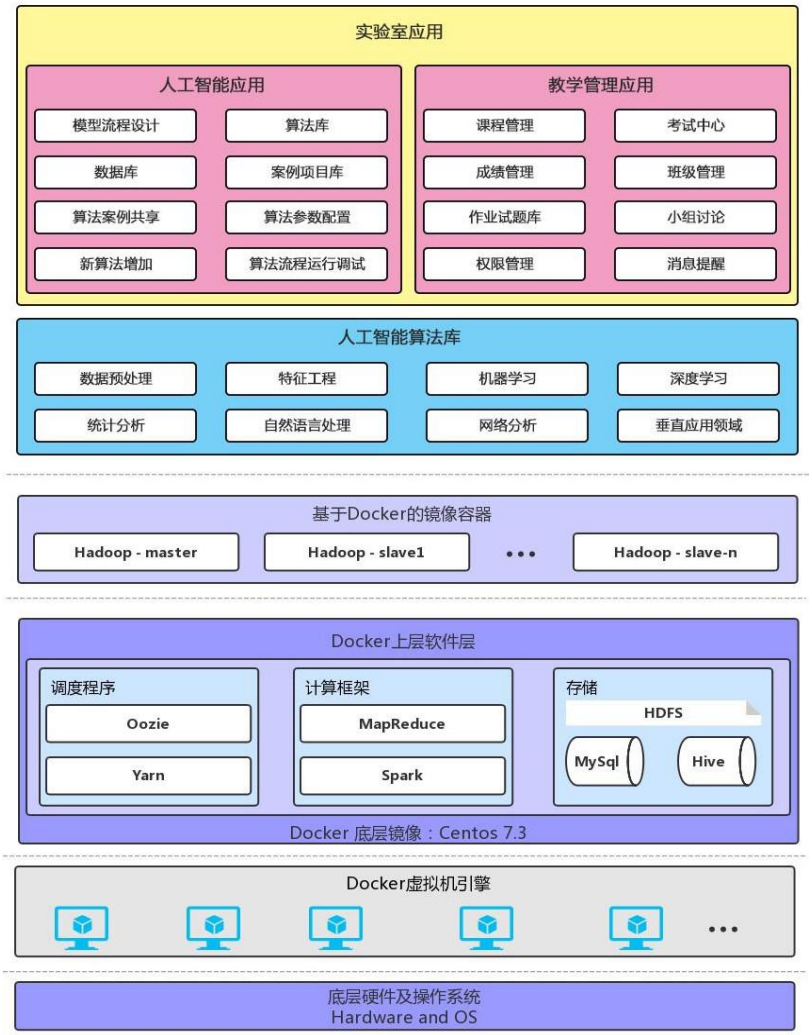
四、人工智能建设内容

针对高校人工智能教学管理、项目实践、科研服务、创新孵化等实际应用场景，人工智能实验室提供稳定、可靠、高效的软硬一体化人工智能教学科研环境，以及完善的课程体系、项目数据和视频、多样化的教学手段和丰富的实战案例。本项目主要包括几个方面的改造建设任务，分别描述如下。

（一）人工智能实验室规划

根据现有国内外高校开展人工智能课程的建设经验以及现有的深度学习及人工智能技术就业市场需求而制定的，让学生具备研究人工智能科学与工程领域问题能力，以及解决人工智能、深度学习应用问题的能力，系统地掌握信息技术、数据分析、机器学习技术、信息处理等基本理论、知识和技能，并且能够根据所学的实训知识与方法技术从事实际的机器学习及人工智能研究工作，具备创新能力和实际操作能力。

（1）人工智能实验室技术规划框架



（图：实验室技术框架）

（2）人工智能实验室技术优势

基于广泛认可的 Hadoop、Spark 分布式文件系统，可有效处理超大规模数据集，具有稳定性高、可扩展性强等优点。

结合 Spark 框架进行分布式数据预处理和算法实现，提供准实时和实时建模的能力， 大大提升了数据的时效性价值。

创新的图形用户界面交互 Free Coding 模式简化建模方式，降低海量数据挖掘的成本和对建模人员编程能力的门槛，使更多学校用户能够享受大数据算法预测模型带来的无限价值。

支持 Python 语言进行交互操作，有助于数据科学家的二次开发。

采用B/S 结构，可以让学生在任何时间和地点做实验，不受环境和时间的限制。老师也可以在任何时间和地点进行实验文档的相关工作。

（二）技术层环境建设

人工智能实验室采用云计算 OpenStack 技术和 Docker 技术，实现硬件资源集中调度和管理，以大数据 Hadoop、Spark 及其生态组件为核心构建软件系统，支持更多的机器学习、深度学习及人工智能技术高级特性，经过专业的性能优化和软硬件集成测试，保证平台的高性能与稳定可靠运行。

（1）人工智能实验室由四部分系统组成

分布式计算系统：是人工智能实验室的资源管理系统；采用 OpenStack 技术和 Docker 技术，实现了硬件资源集中调度和管理。

AI 工程平台:是人工智能实验室的数据处理平台,实现数据解析、模型训练、模型预测、异常值处理、模型评估、模型发布，为人工智能实验室提供平台支撑。

人工智能训练系统：是人工智能实验室的核心训练系统；由课程管理系统、资源管理系统、实验报告考评系统、考试系统组成，为实验教学提供技术支撑。

人工智能训练包：是人工智能实验室的课程实验内容部分，由基础算法训练集、深度学习算法训练集、行业应用训练集组成。

分布式计算系统：分布式计算系统通过对硬件设施进行虚拟化处理，形成虚拟层面的资源池系统，该资源池系统可按需为每一套应用系统提供基础 IT 资源——计算能力、存储能力和网络功能，快速适应不断变化的业务需求，实现“弹性”资源分配能力。客户通过统一的 Web 界面，可实现对整个数据中心集中管理，包括虚拟机、资源池、数据中心等，从而为用户提供可靠、优质的计算服务。

分布式计算系统的特点

自动化远程部署：通过分布式计算系统自动化部署工具，可以实现在高校现有网络环境中，对裸机进行自动化安装，真正做到即买即用。

安全可靠：云环境支持 VLAN\VXLAN\GRE 网络隔离、软防火墙、平台各关键组件采用 NRW 模型， 保证系统的高可用。对 Linux 的云主机提供 ssh-key 登录方式的密钥，使得访问更加安全可靠；对关键的操作进行记录，让操作行为有据可查。

易扩展：通过分布式计算系统自动化部署工具，可以自动化管理，实时监控个节点，支持弹性伸缩。

全方位监控：全方位实时秒级监控，包括 CPU、内存、磁盘、网络等资源，并针对个性化需求自定义报警阀值，及报警处理，保证您应用的性能及可用性。

异构支持：同时支持 KVM\Hyper-V\VMware\Xen 等异构虚拟化技术、存储异构可以支持本地存储\ 分布式存储\FC SAN\IP SAN 等多种设备。

分布式存储：提供无 Vendor-Lock-in 的存储方案，支持 x86 架构，从TB 到PB级的扩展，没有单点故障，多数据副本，自动管理，自动修复，数据分布均衡，并行化程度高。

SDN：支持纯软件的 SDN 解决方案，上层网络资源可以由用户自由定义，同时也支持各个硬件 SDN 交换机厂商提供的针对于其 SDN 产品，可根据需求定制开发 OpenStack 网络驱动。

开放创新：紧跟 OpenStack 社区技术发展，提供原生 OpenStack API、扩展功能提供标准 REST API，用户可以自行扩展。

自动化运维：利用 Pacemaker+Haproxy 的架构，实现集群的高可用和自动化管理，当集群中某个节点某功能出现故障，可以自动化该功能进行恢复，同时通知 haproxy 将请求分发到功能正常的节点上。

方便使用：利用分布式计算系统管理系统，可以快速创建云主机、私有网络、DHCP、路由器、防火墙等功能。

分布式计算系统功能框架及介绍：分布式计算系统，由计算、镜像、块存储、网络和安全等几个主要的组件组合起来完成具体工作。依托 OpenStack、Docker 技术的云支持几乎所有类型的云环境，目标是提供实施简单、可大规模扩展、丰富、标准统一的云计算管理平台。

计算模块：计算模块主要提供云主机功能。而云主机提供了整个云平台中最基础的功能，即虚拟服务器从创建到销毁的全生命周期维护。此模块通过利用虚拟化技术，可将大批服务器硬件资源池化，用户仅需点击鼠标，选择期望的硬件配置、操作系统类型和网络配置等信息，即可在短时间内按需获得任意数量的云主机，模块支持云主机硬件配置在线升级、云主机热迁移、重启、暂停、创建快照等多种功能。

（图 3-2）

（图 3-3）

镜像模块：镜像功能模块是一套虚拟机镜像查找及检索系统，支持多种虚拟机镜像格式（AKI、AMI、ARI、ISO、QCOW2、Raw、VDI、VHD、VMDK），有创建上传镜像、删除镜像、编辑镜像基本信息的功能。镜像功能模块用于为云主机提供一种快速部署的方式，通常包含了操作系统、应用程序及相关服务，用户通过此模块可以“秒级”创建一台云主机。

镜像，即 Image，通常分为两类：即公有镜像和私有镜像。

公有镜像：即整个云平台的基础镜像包，可结合客户的场景需求，由我司预安装在平台上，供所有用户直接调用。镜像类型支持业界主流的、运行于 x86 架构的操作系统，具体如下表所示。

类型	版本
windows	Win7/win8/win2008R2/win2012 等含 32 位和 64 位
Linux	RHEL（CentOS）5.6 以上/6.x/7x 等含 32 位和 64 位

（表 3-1：镜像配置）

私有镜像：即由用户自定义的镜像包，用户可将自己定制化的云主机转换为私有镜像，从而实现批量复制/创建的功能，有镜像仅供创建该镜像的用户使用，也可以转化为公有镜像。

（图 3-4：镜像）

块存储模块：块存储模块为运行实例提供稳定的数据块存储服务，即云硬盘服务。它的插件驱动架构有利于块设备的创建和管理，如创建卷、删除卷，在实例上挂载和卸载卷。它们独立于云主机的生命周期而存在，可挂载到任意运行中的云主机上，确保单台云主机故障时，数据不丢失，并具备基于云硬盘的快照创建、备份和快照回滚等功能。

（图 3-5：存储模块）

网络模块：网络模块提供云计算的网络虚拟化技术，为云平台其他服务提供网络连接服务。为用户提供接口，可以定义 Network、Subnet、Router，配置 DHCP、DNS、负载均衡、L3 服务，网络支持 GRE、VLAN。插件架构支持许多主流的网络厂家和技术，如 OpenvSwitch。

网络功能模块基于先进的 SDN（Software Defined Network）技术，为用户提供按需构建网络的功能，包括虚拟交换机、虚拟路由器、公网 IP、VPN 和负载均衡等子功能，通过实时同步的网络拓扑展现功能，可有效提高用户构建复杂网络的效率。

安全模块：安全模块通过在计算模块中添加扩展实现，基于传统的包过滤型防火墙技术，可为用户的云主机提供细颗粒度的安全防护策略，支持TCP/UDP/ICMP 等多种协议，支持自定义来源IP和端口范围等规则，支持用户针对不同类型云主机加载不同级别安全策略的功能。

（图 3-6：安全模块）

（图 3-7：安全模块）

（2）AI 工程平台

2.1 工程化管理平台模块

工程化管理平台实现了对各数据建模整个生命周期的可视化和模块化管理，并以友好的用户界面和高级的技术特性，整合用户管理、任务管理、数据管理和模型管理等业务级管理任务。

数据建模工程界面：友好且实用性极强的图形用户界面交互 Free Coding 模式；

数据的工程化上传、存储、加载和管理；

模型的工程化创建，调优，存储，加载和管理；

展示性的 AI 工程模型仓库；

数据的快速存储和加载功能；

AI 工程平台的数据存储和加载功能模块基于 Hadoop/Spark 集群，通过分布式文件系统，HDFS 的数据接口，提供数据整合和数据质量管理等技术，支持海量数据的快速存储和加载。

海量数据的快速存储:基于分布式文件系统 HDFS 的集群分布式数据存储和列表显示，支持 Hadoop/Spark 的访问接口；

海量数据的数据质量加速器:交互式数据质量管理操作，包括数据创建、数据拆分和数据整合；海量数据的分布式加载、数据上传和导入。

2.2 数据预处理和统计分析

AI工程平台集合了众多常用的数据处理和统计分析技术，通过交互式和可视化的工具， 实现数据整理、变量分析、特征刻画、相关性探索和数据可视化等，支持对数据快速分析和整体把握。

海量数据的数据整理；

丰富的数据预处理模板:数据清洗（缺失值处理）、抽取转化、标准化、分箱（离散化）、独热编码（One-Hot Encoding）等；

海量数据的统计分析；

数据的变量分析：数据的统计特征，核密度估计，维度相关性分析；

数据的特征刻画： 线性回归、统计量计算、分层抽样、方差分析、显著性检验（假设检验）、时间序列等；

数据可视化

各个分量数据的分布

多分量数据的相关性可视化

全量数据的描述性建模

AI 工程平台集合众多主流的机器学习算法，结合Hadoop/Spark平台的分布式能力，支持基于海量数据集的全量数据描述性建模，并且提供菜单式参数调优界面，实现了企业级AI模型生产和分析。

主流的机器学习算法的描述性建模:

分类：深度学习、随机森林、朴素贝叶斯模型、广义线性；

聚类：K-means

回归：广义线性模型，梯度提升模型；

降维：主成分分析，广义低阶模型；

时间序列：ARIMA, GARCH；

探索性数据建模策略：

数据变量的特征选择：信息增益，决策树；

建模数据的交叉验证：N折交叉验证；

菜单式参数调优选择：各模型各参数的提示性参数设置；

描述性建模的模型评价；

ROC 曲线和AUC值；

准确率、精准率、召回率、F1-measure；

多种评判准则下的预测数结果矩阵；

预测性建模及模型评判；

AI 工程平台基于海量数据的描述性探究建模结果，通过对模型和数据的再处理，得到数据的独立化预测性模型，实现了对测试数据的一键式预测，并提供模型评判和预测报告。

预测性建模自动化和独立化；

预测模型的训练数据处理和模型建立的程式化创建和存储。

训练数据所得的预测模型独立分装为分类器

模型结果的显示化表达

测试数据和标签的显示化展示

测试数据的各项结果和测试标准的展示

模型报告和评判

模型测试结果评判和报告

多个建模结果的比较

自主编程及特定场景开发

AI 工程平台集成了多种编程环境，支持用户的自主开发，以及特定场景的多环境编程， 实现针对特定客户的系列业务开发。

集成 Scala/Python 编程环境，用户自主编程开发

特定场景开发：特定场景的模型开发和模型仓库存储

（三）人工智能训练系统

人工智能实验室共包含 7个模块，分别是实验室、数据库、算法库、案例库、考试中心、课程管理和班级管理。

其中考试中心包含 3 方面内容，分别是试题库、试卷库、作业库；

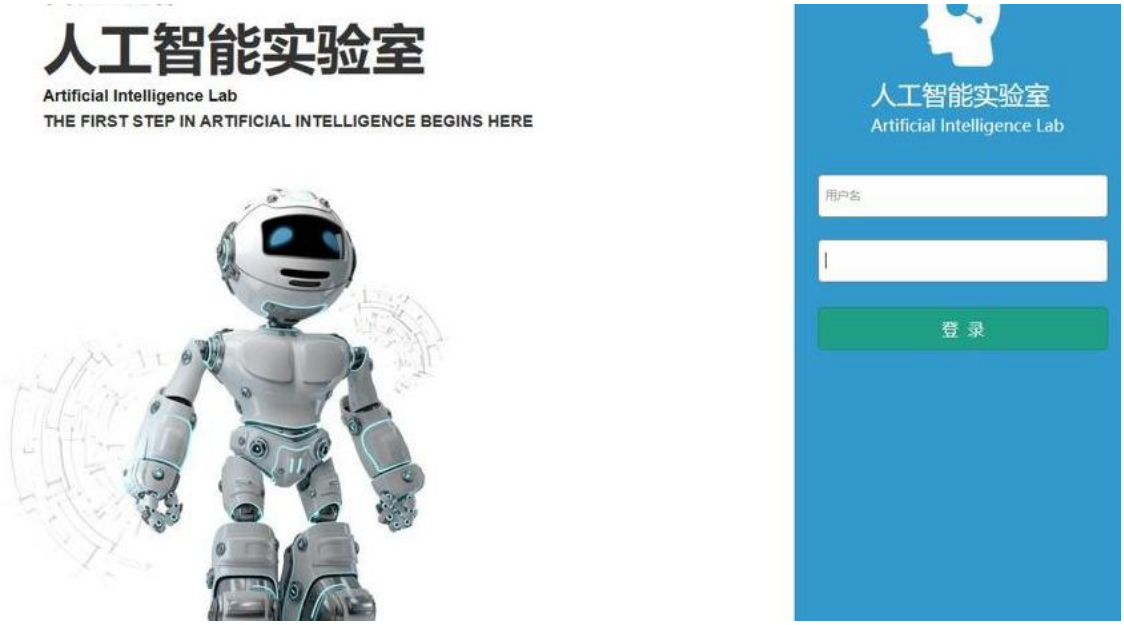
课程管理包含 2 方面内容，分别是创建课程、课程管理；

班级管理包含 4 方面内容，分别是我的班级、作业审阅、试卷审阅、成绩管理。

人工智能训练系统功能介绍；

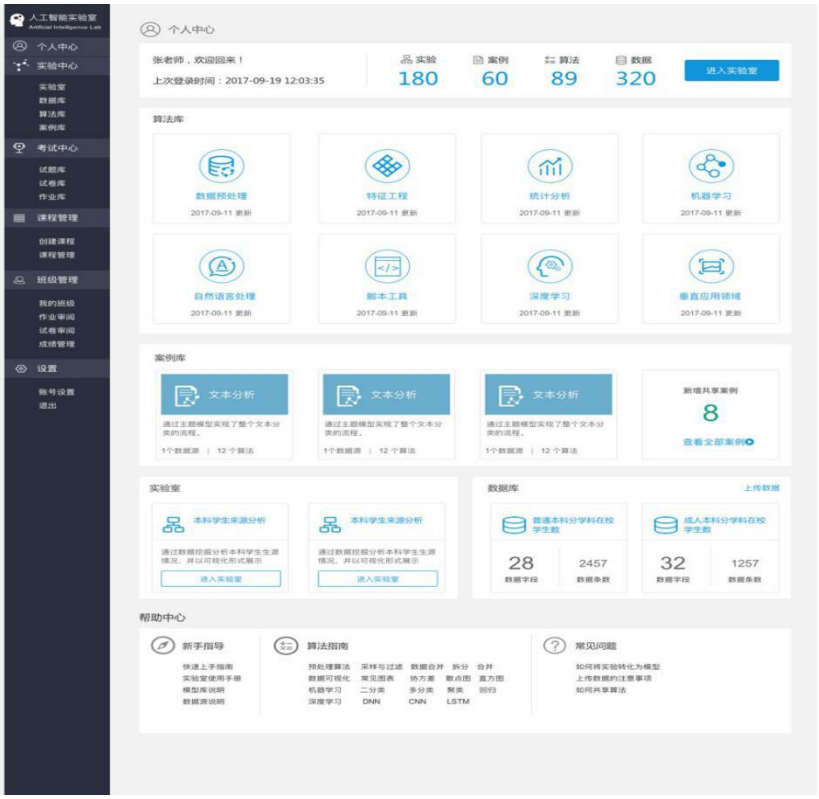
登录

在浏览器打开登录界面，输入用户名和密码进行登录。界面如下：



（图 4-1）

（1）个人中心



（图 4-2）

教师进入系统后的首页界面，提供 5 项内容，包含算法库、案例库、实验室、数据库、帮助中心。通过顶端全局计数统计可了解实验总体情况，点击【进入实验室】可直接快速的进入上次打开的实验画布。

算法共提供 9 个分类，包含数据预处理、特征工程、统计分析、机器学习、自然语言处理、脚本工具、深度学习和垂直应用领域。鼠标移上，显示此分类下具体的算法名称。

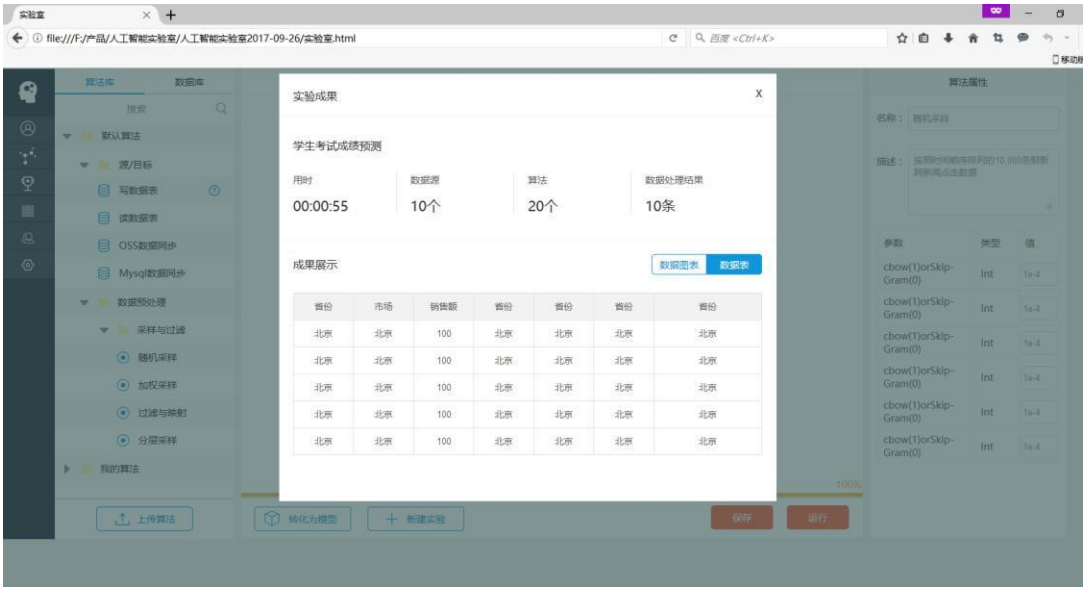
案例库提供 3 方面内容，包含最新的案例、新增共享案例数量、查看全部案例入口；其

中案例信息包含案例描述、数据源应用数量、算法应用数量 3 部分内容。

实验室展示最新的 2 个实验，实验信息提供 3 部分内容，包含实验名称、实验描述和实

验室入口；数据库展示最新使用的 2 个数据，数据信息提供 4 部分内容，包含数据源名称、数据字段数量、数据条数、上传数据入口。

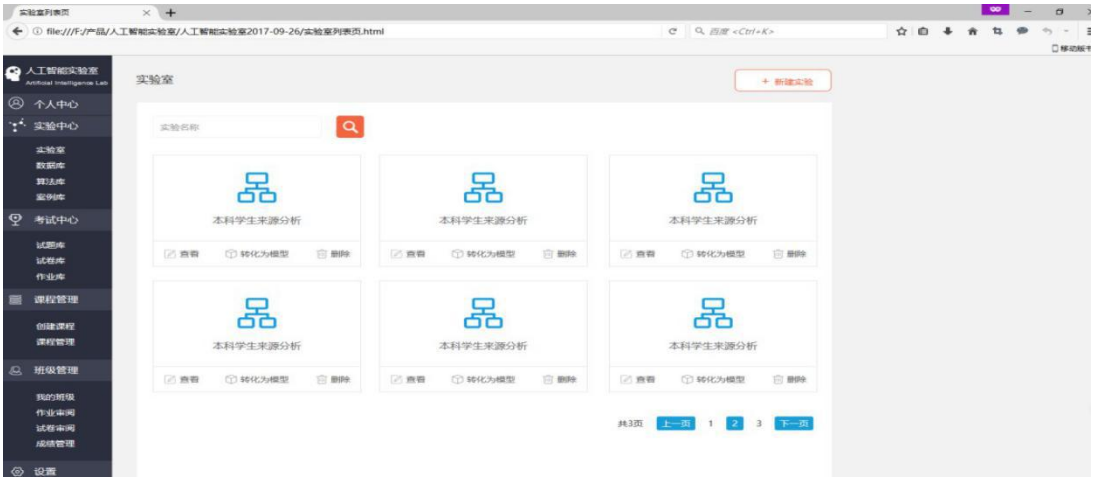
帮助中心是为了用户能自助解决使用中的问题，提供 3 部分内容，包含新手指导、算法指南和常见问题。



（图 4-3）

（2）实验画布

1）实验列表

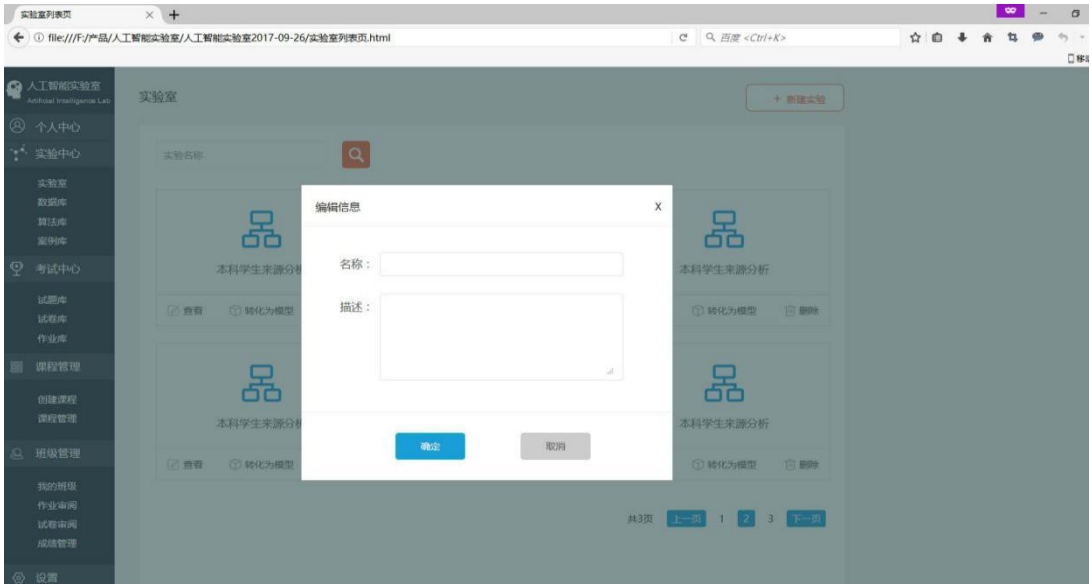


（图 4-4）

点击左侧导航“实验室”，进入实验列表界面，提供新建实验、搜索实验、查看实验、转化为案例、删除 5 种操作。其中，点击查看进入实验画布界面；点击转化为案例进入将实验算法流程转化为案

例界面；点击新建实验显示设置实验信息弹窗。

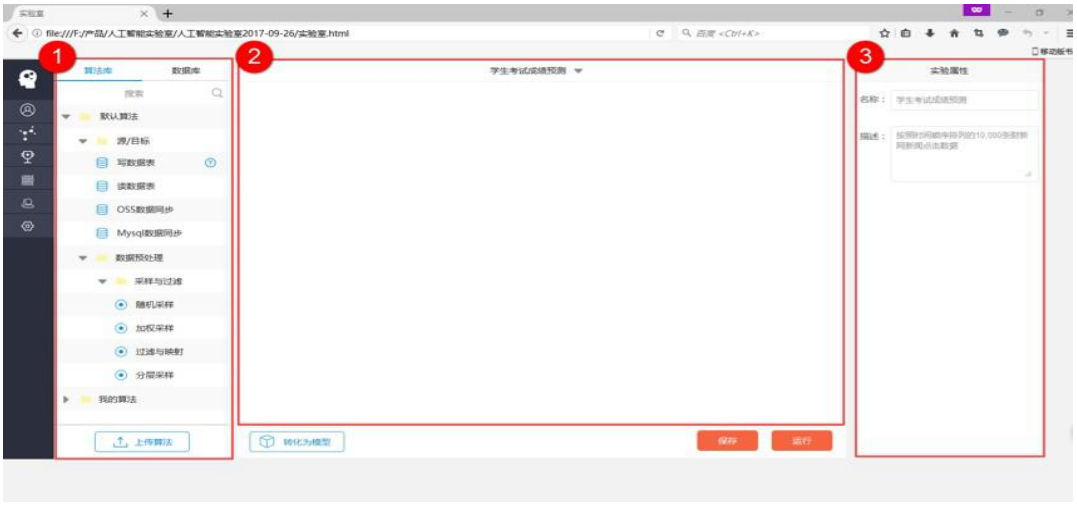
2)新建实验



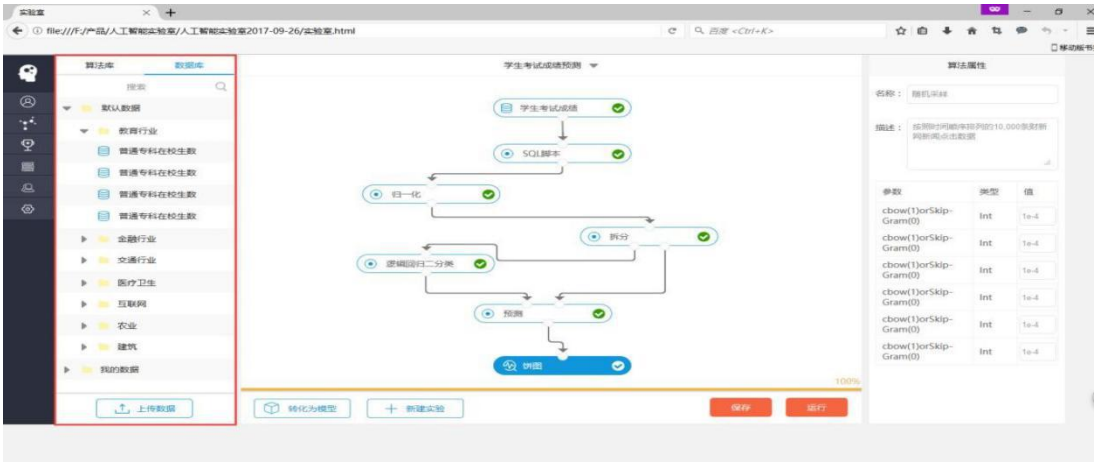
（图 4-5）

在实验列表界面，点击新建实验显示设置实验信息的弹窗，设置内容包含实验名称和实验描述 2 项。点击确定，即可进入实验画布界面。

画布介绍



（图 4-6）

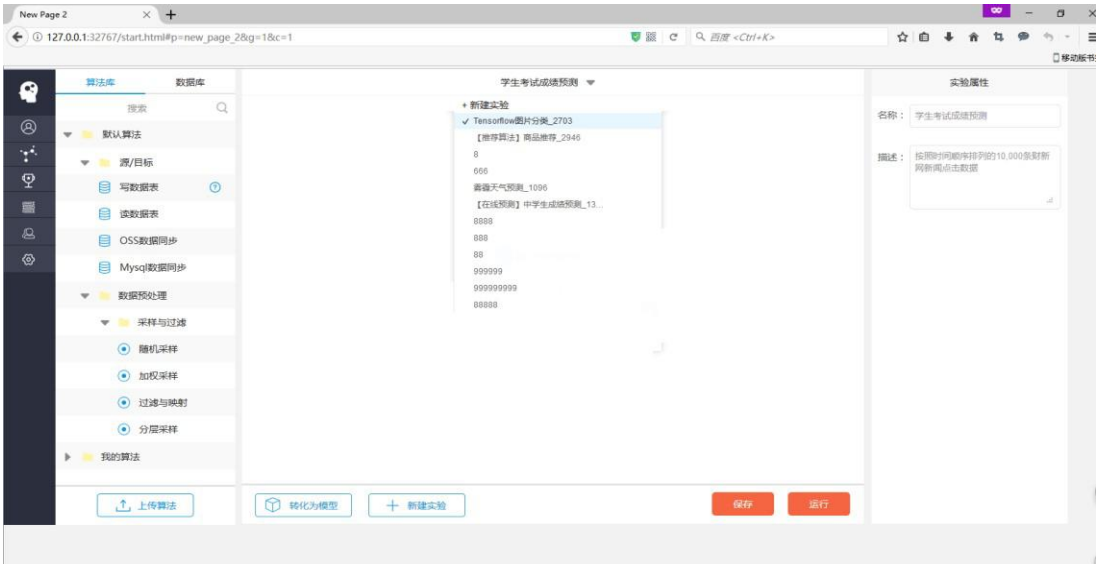


（图 4-7）

在实验列表界面，点击查看可进入实验画布界面；在个人中心界面也提供有快速进入实验画布界面的入口，即“进入实验室”按钮。

实验画布界面分为左中右 3 部分内容，其中左侧部分为算法列表和数据列表；中间部分为操作面板，左侧算法和数据均可自由拖拽至操作面板；右侧部分为属性设置，可设置实验属性、算法属性和数据属性；当没有选中任何算法和数据时，右侧显示设置实验属性；当选中某算法时，右侧可设置此算法的属性；当选中某数据时，右侧可设置此数据的属性。

切换实验

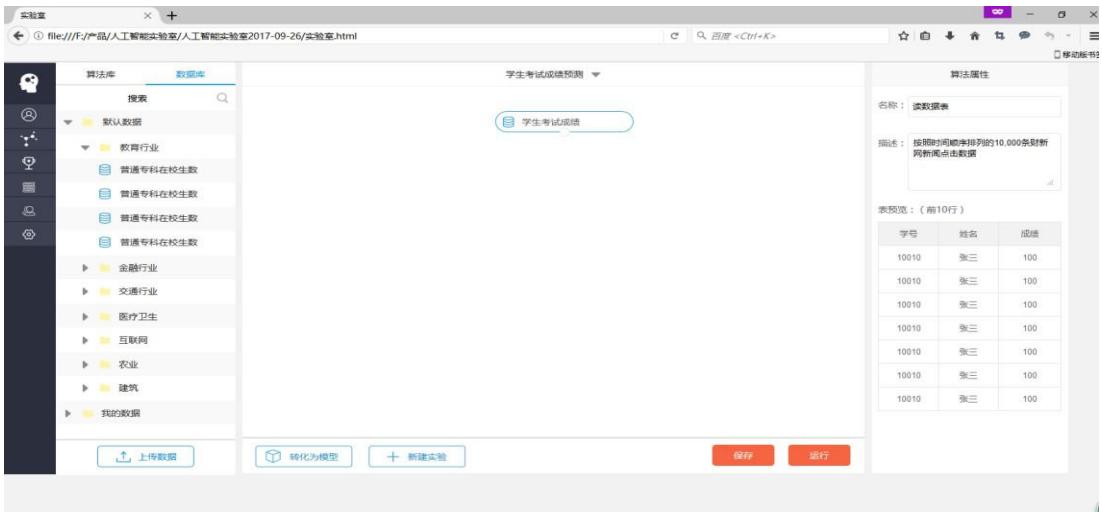


（图 4-8）

在实验画布界面中间部分，点击实验名称可显示实验列表浮层。在浮层里点击其他实验名称，可快速切换至对应实验；点击新建实验，显示新建实验弹窗，快速创建新的实验。

选择数据源

实验画布界面，左侧默认显示算法列表，点击数据库切换至数据列表，数据列表分 2。

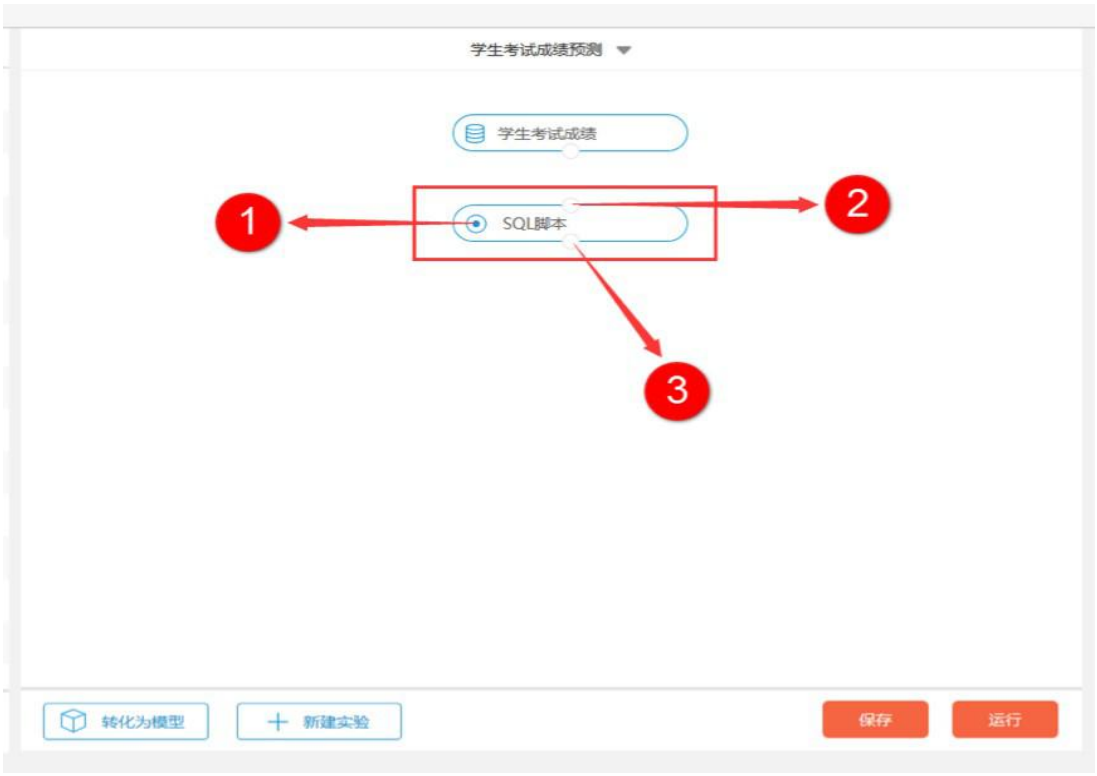


类 3 级，包含默认数据和我的数据 2 类，默认数据为系统提供，我的数据为用户自己上传。拖拽需要的数据至中间面板，界面如下：

（图 4-9）

选择算法

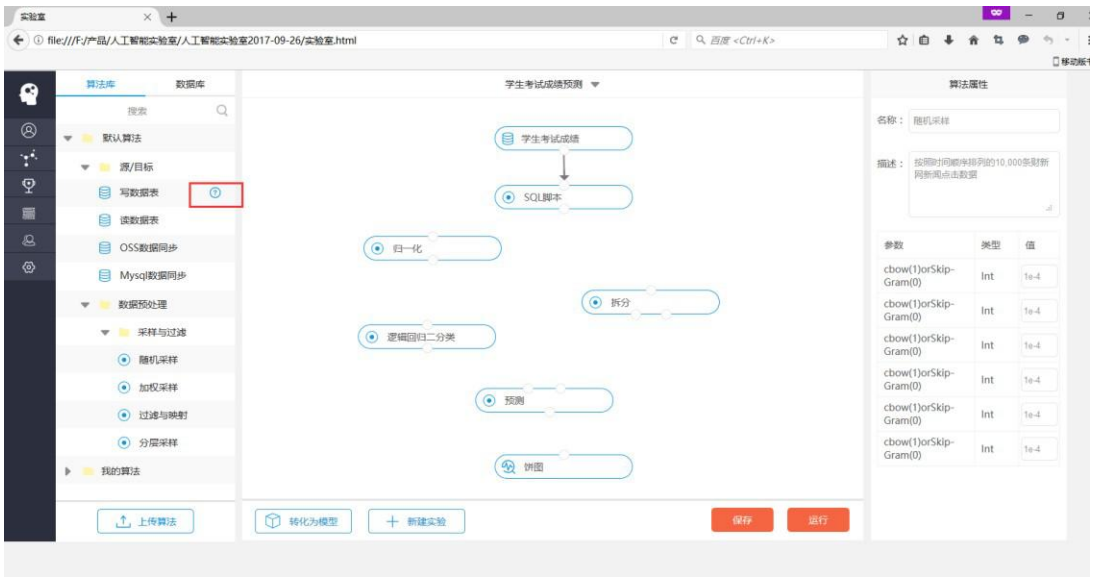
同样也支持拖拽左侧的算法至中间操作面板，算法和数据没有硬性的前后拖拽要求，依用户习惯拖拽。



（图 4-10）

算法拖拽至操作面板后，显示内容包含 3 部分，即算法图标（图中标识 1）、输入点（图中标识 2）、输出点（图中标识 3）。不同分类的算法对应的图标不同，且不同算法的输入点数量和输出点数量均不相同。不同算法的输入点和输出点可自由连线。

算法说明

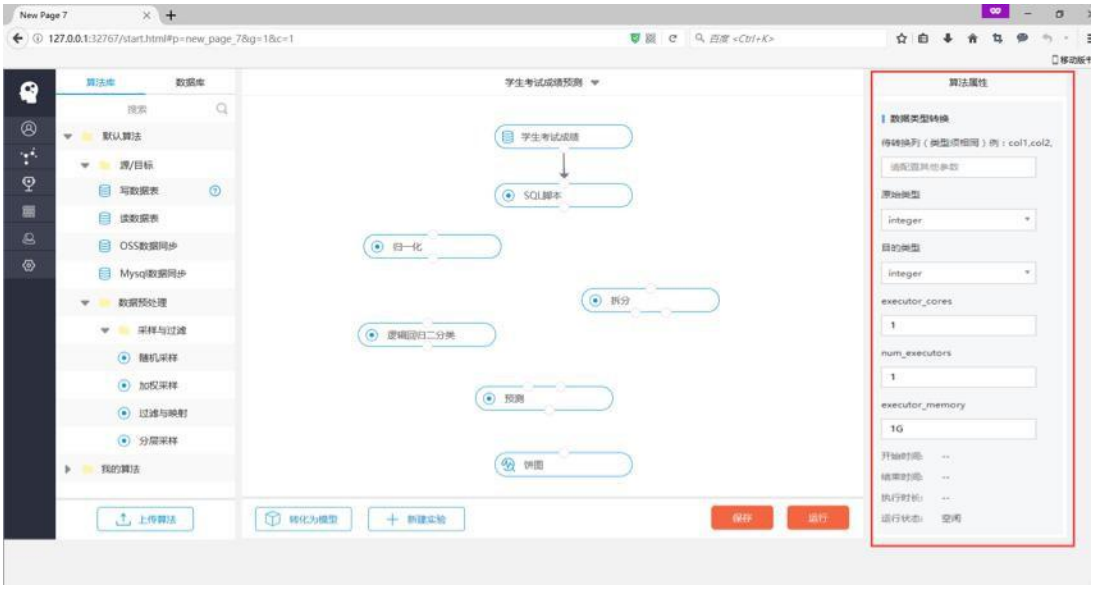


（图 4-11）

实验画布左侧算法列表部分，鼠标移上某算法名称，均可显示蓝色提示图标， 点击图标进入算法详细说明界面，辅助用户进行算法属性设置。

算法配置

点击选中某算法，在实验画布右侧可进行算法属性设置。不同算法对应的设置项及内容不同。设置正确与否，点击运行均会有对应提示。



（图 4-12）

（3）运行并调试算法

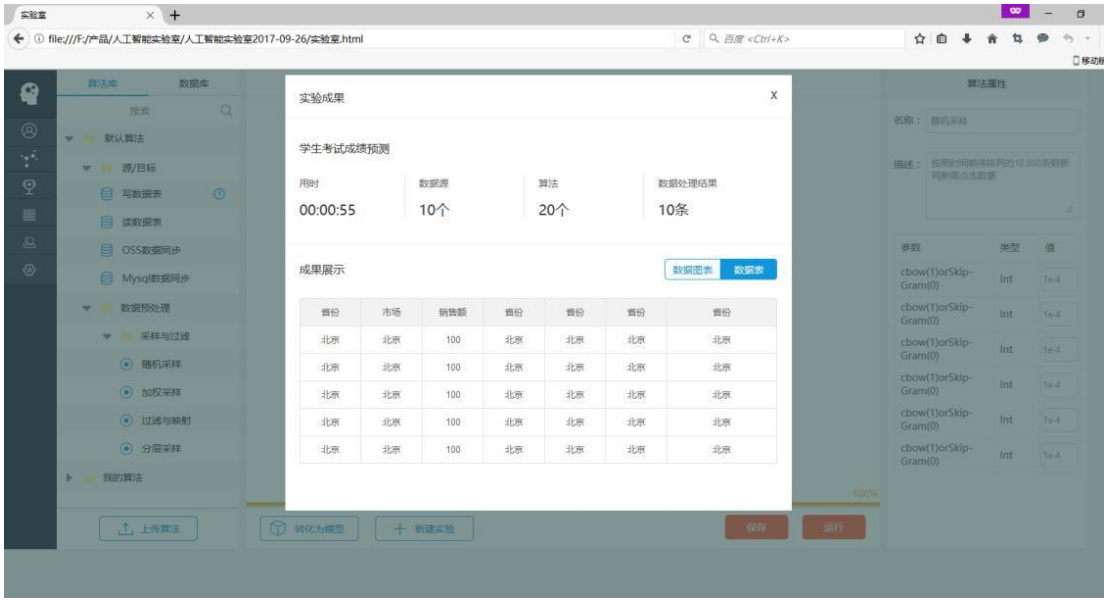
拖拽算法和数据完毕后，算法和数据间可自由连线，连线完毕，点击运行，依次运行实验组件，运行成功即可查看实验成果。

目前正在运行哪一算法，有蓝色动态箭头显示（即图中标识 1），箭头的引出方为正在运行的算法。算法属性设置正确，则显示绿色的对号图标（即图中标识 2）；算法属性设置错误，则显示红色的叹号图标（即图中标识 3）及错误提示（即图中标识 4），可根据错误提示进行算法属性修改；下方同时显示当前运

行进度（即图中标识 5）。

3.1 结果展示

（图 4-14）



实验运行成功后，可视化算法会呈蓝色，点击可视化算法，可查看实验成果。

（图 4-15）

实验成果提供 3 项内容，包含实验数据（用时、数据源使用量、算法使用量、数据处理结果条数）、数据图表、数据处理结果表。数据图表和数据处理结果表之间可自由切换，数据图表可根据数据切换图表类型，数据处理结果以图表的形式展示。

3.2 算法库

算法总览

算法列表

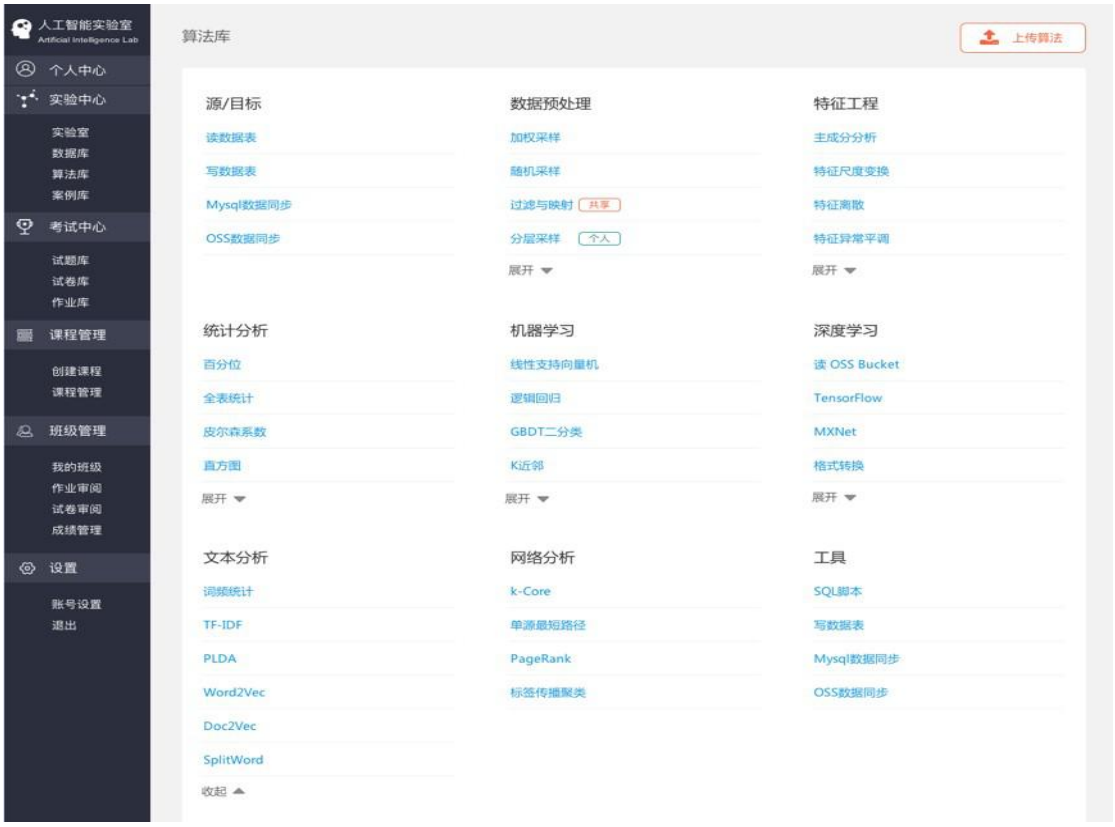
点击左侧导航进入算法库列表页，可以查看全部可使用的算法名称，点击进入算法详细说明页面。

算法名称有【共享】标签的则为用户自主上传的算法并可以共享其他人使用； 算法名称有【个人】标签的则为用户自主上传仅能自己使用；

点击展开可以看到该分类下其他的算法。

算法介绍及参数说明：点击进入算法详细说明页，具体的介绍该算法的使用文档。算法介绍：详细介绍该算法使用的场景，以及注意事项等；

参数说明：介绍该算法可以配置的参数，包括参数名称、参数描述、可选项和默认值等。



（图 4-18）

若是自己上传的算法点击右上角的【共享】按钮可以共享给其他用户使用。

（图 4-19）

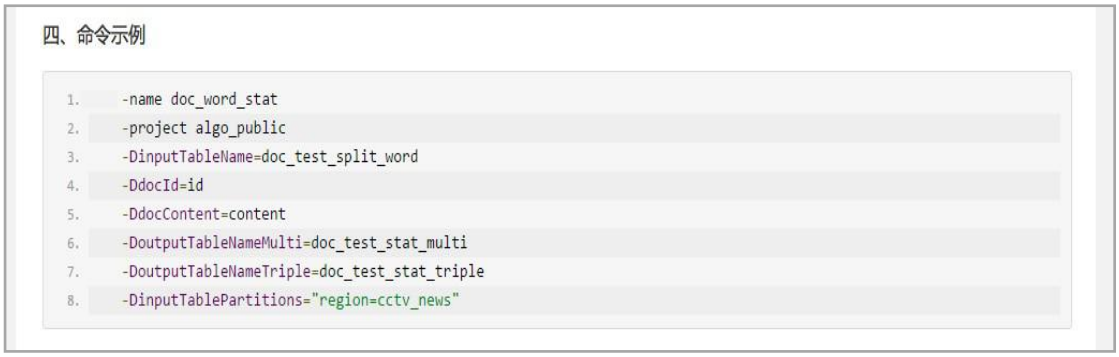
■ 参数设置

说明如何通过实验画布设置算法的参数以及注意说明。

（图 4-20）

■ 命令实例

介绍算法中核心命令代码，方便用户了解算法实现原理



■ 实例

实例部分的测试数据展示该算法需要的数据源结构，并举例说明； 运行

命令展示数据与算法如何对接和处理；

五、实例

1.测试数据

id	y	f0	f1	f2	f3	f4	f5	f6	f7
1	-1	-0.294118	0.487437	0.180328	-0.292929	-1	0.00149028	-0.53117	-0.0333333
2	+1	-0.882353	-0.145729	0.0819672	-0.414141	-1	-0.207153	-0.766866	-0.666667
3	-1	-0.0588235	0.839196	0.0491803	-1	-1	-0.305514	-0.492741	-0.633333
4	+1	-0.882353	-0.105528	0.0819672	-0.535354	-0.777778	-0.162444	-0.923997	-1
5	-1	-1	0.376884	-0.344262	-0.292929	-0.602837	0.28465	0.887276	-0.6

2.运行命令

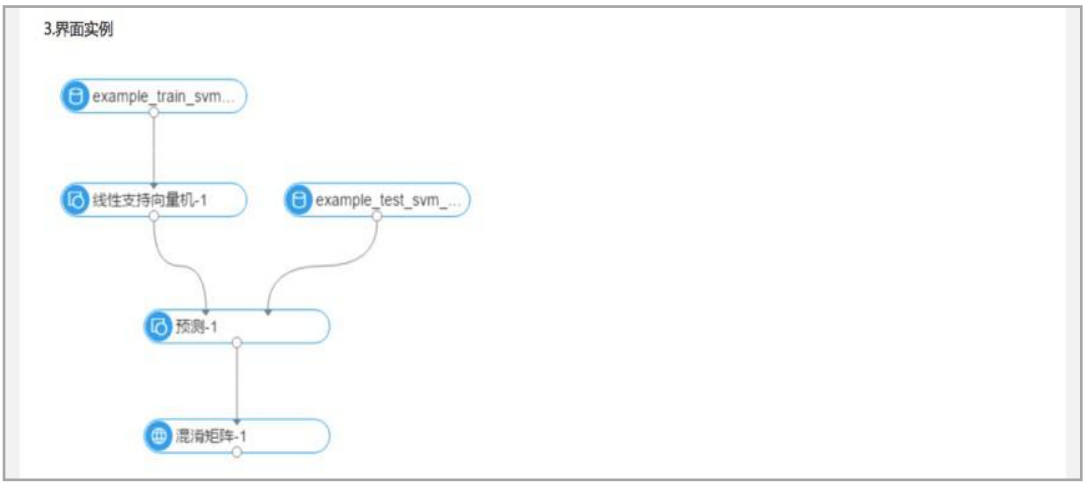
```
drop table if exists enum_feature_selection_test_input_enum_value_output;
drop table if exists enum_feature_selection_test_input_cnt_output;
drop table if exists enum_feature_selection_test_input_value_output;
PAI -name enum_feature_selection -project algo_public -DitemDelimiter=":" -Dlifecycle="28" -
DoutputValueTableName="enum_feature_selection_test_input_value_output" -DkvDelimiter="," -DlabelColName="col_bigint"
-DfeatureColNames="col_double,col_string" -
DoutputEnumValueTableName="enum_feature_selection_test_input_enum_value_output" -DenableSparse="false" -
DinputTableName="enum_feature_selection_test_input" -
DoutputCntTableName="enum_feature_selection_test_input_cnt_output";
```

（图 4-21）

■ 算法界面操作

通过举例说明该算法如何在实验画布中使用并设计算法流程。

■ 运行结果



举例通过该算法计算后的数据结果。

5.运行结果

enum_feature_selection_test_input_cnt_output

```
+------------+------------+------------+------------+
| colname    | colvalue   | labelvalue | cnt        |
+------------+------------+------------+------------+
| col_double | NULL       | 1          | 1          |
| col_double | 0          | 0          | 2          |
| col_double | 0          | 1          | 1          |
| col_double | 1          | 0          | 1          |
| col_double | 1          | 1          | 1          |
| col_string | NULL       | 0          | 1          |
| col_string | 00         | 0          | 1          |
| col_string | 00         | 1          | 1          |
| col_string | 01         | 0          | 1          |
| col_string | 01         | 1          | 2          |
+------------+------------+------------+------------+
```

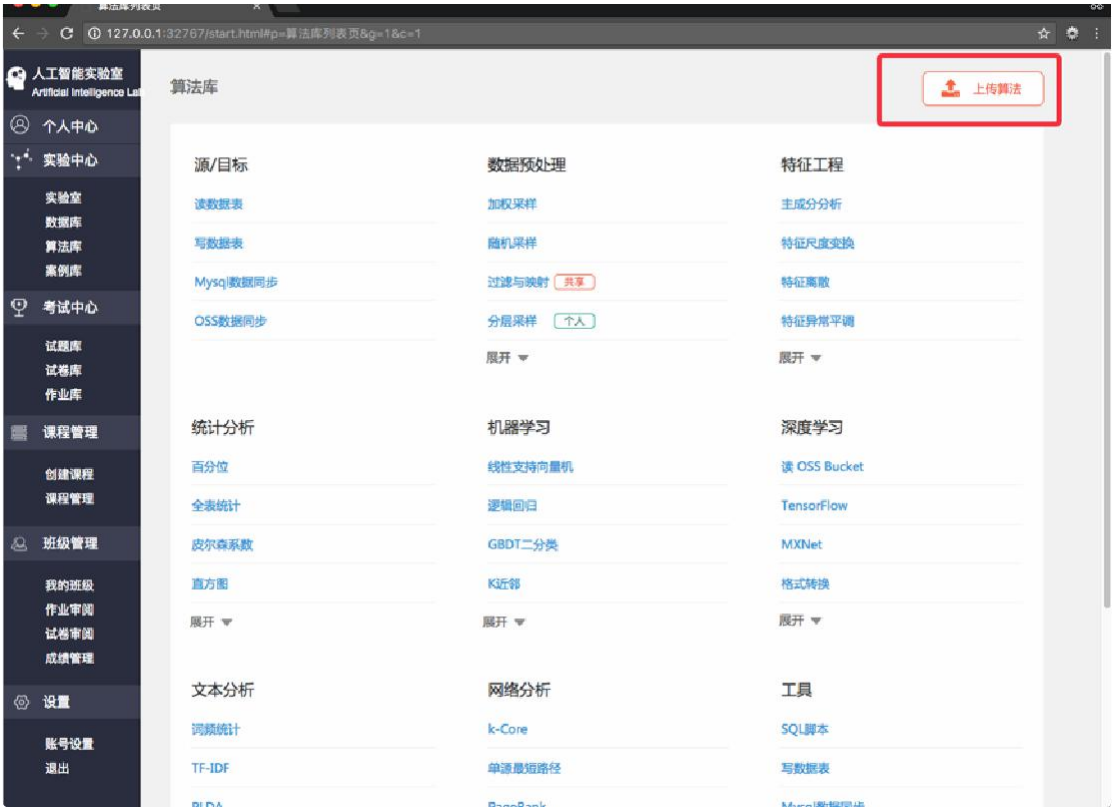
（图 4-23）

■ 界面运行结果

图形化展现数据结果举例。

（图 4-24）

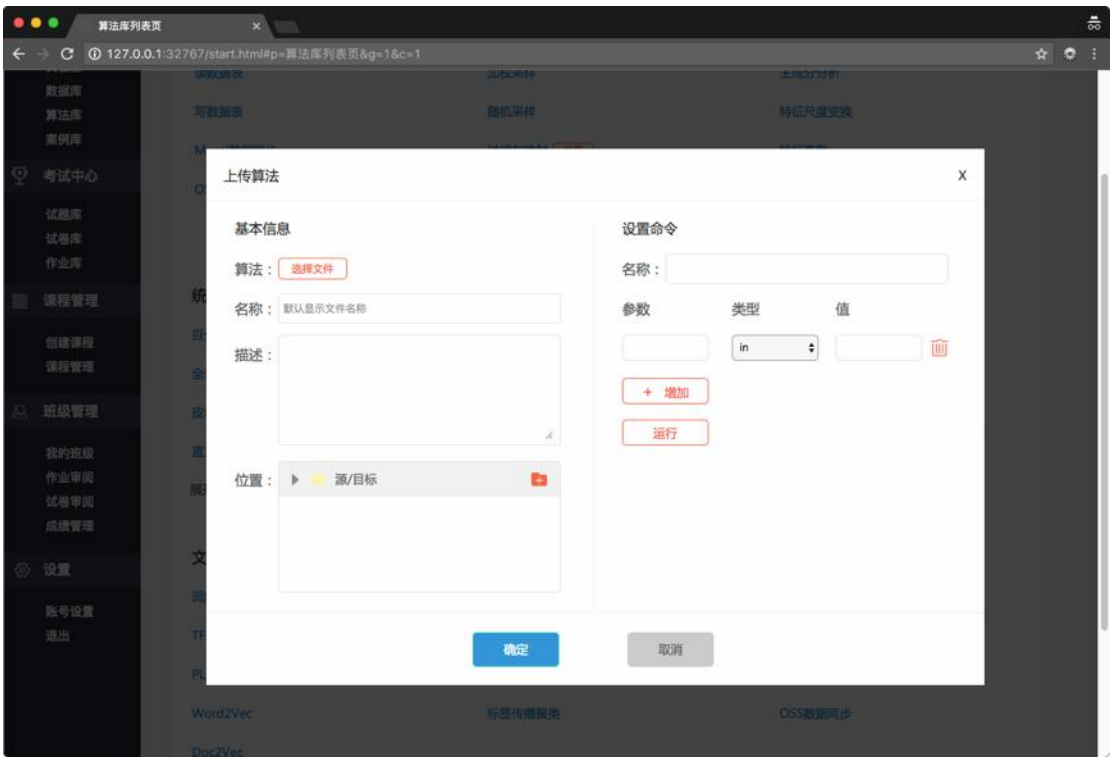
2） 新算法上传



（图 4-25）

通过算法列表页或实验画布页的【上传新算法】可打开算法上传界面。

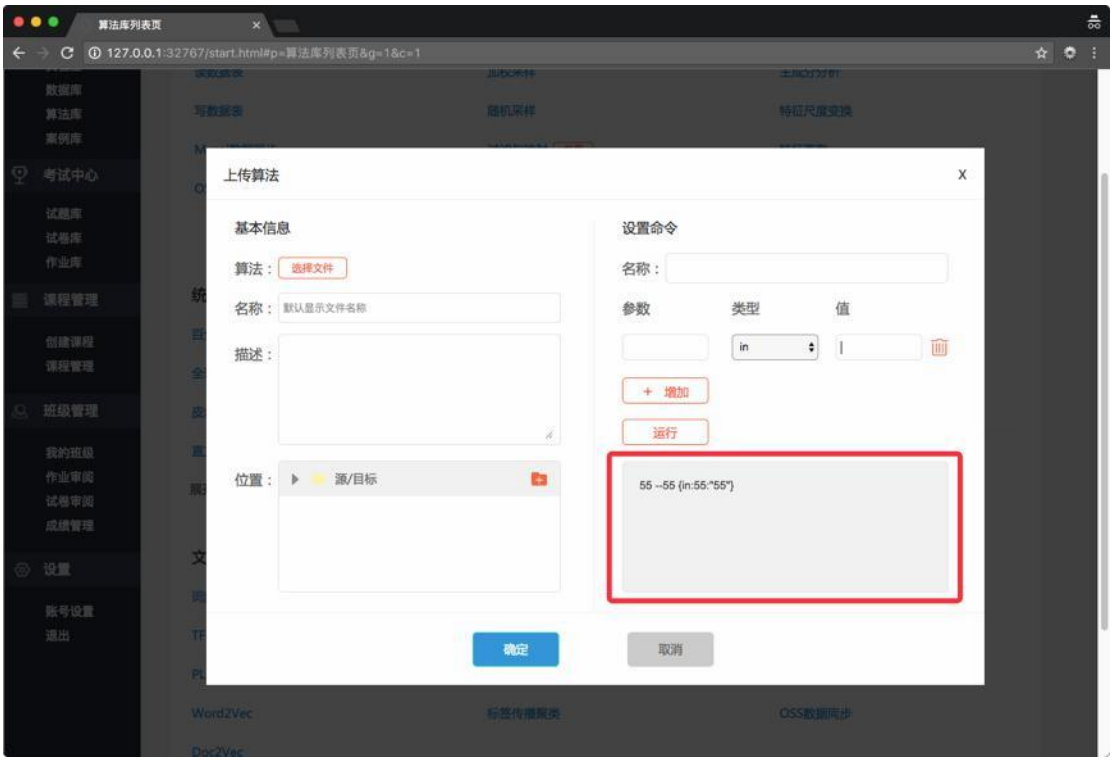
- 上传算法



（图 4-26）

选择本地算法包，填写算法相关内容包括算法名称、描述，选择算法分类； 设置算法参数，包括参数名称、类型、默认值，支持参数新增、删除；

- 算法运行测试



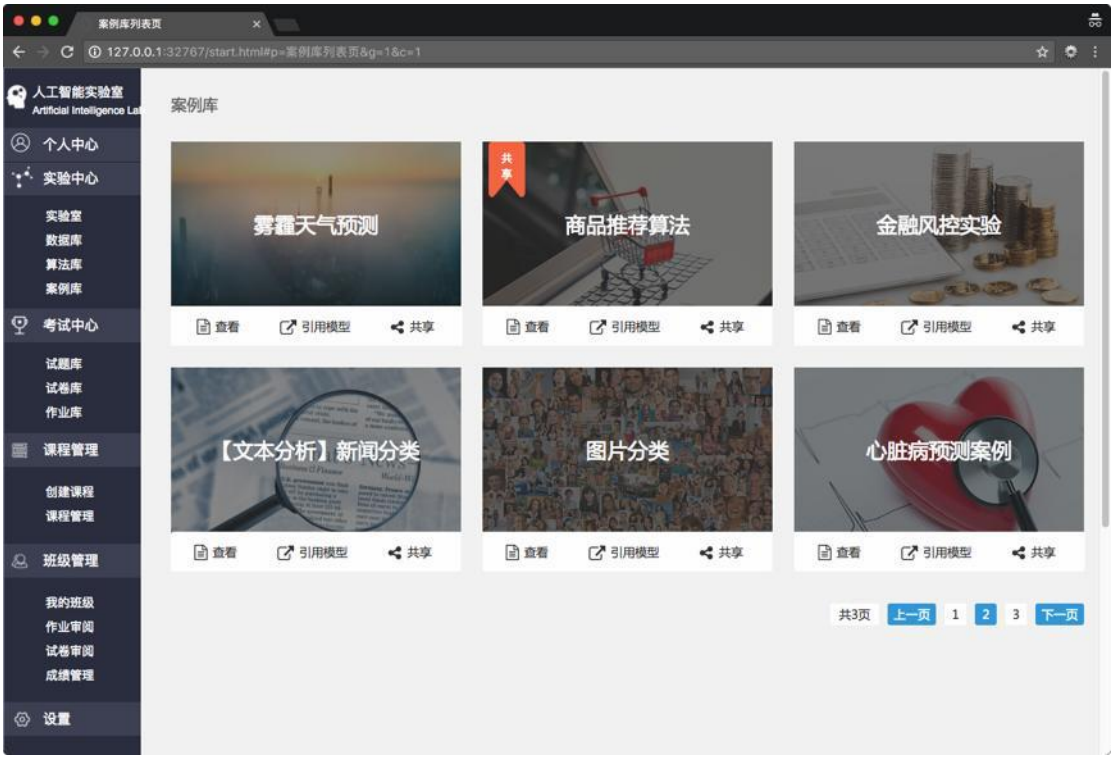
（图 4-27）

点击运行并输出算法结果，测试算法是否成功上传。

成功上传的算法即可在实验画布左侧算法库中显示及使用。

3.2 案例库

案例库列表



（图 4-28）

实验室默认提供经典的人工智能项目案例作为用户学习教学使用，点击右侧导航中的【案例库】即可进入。

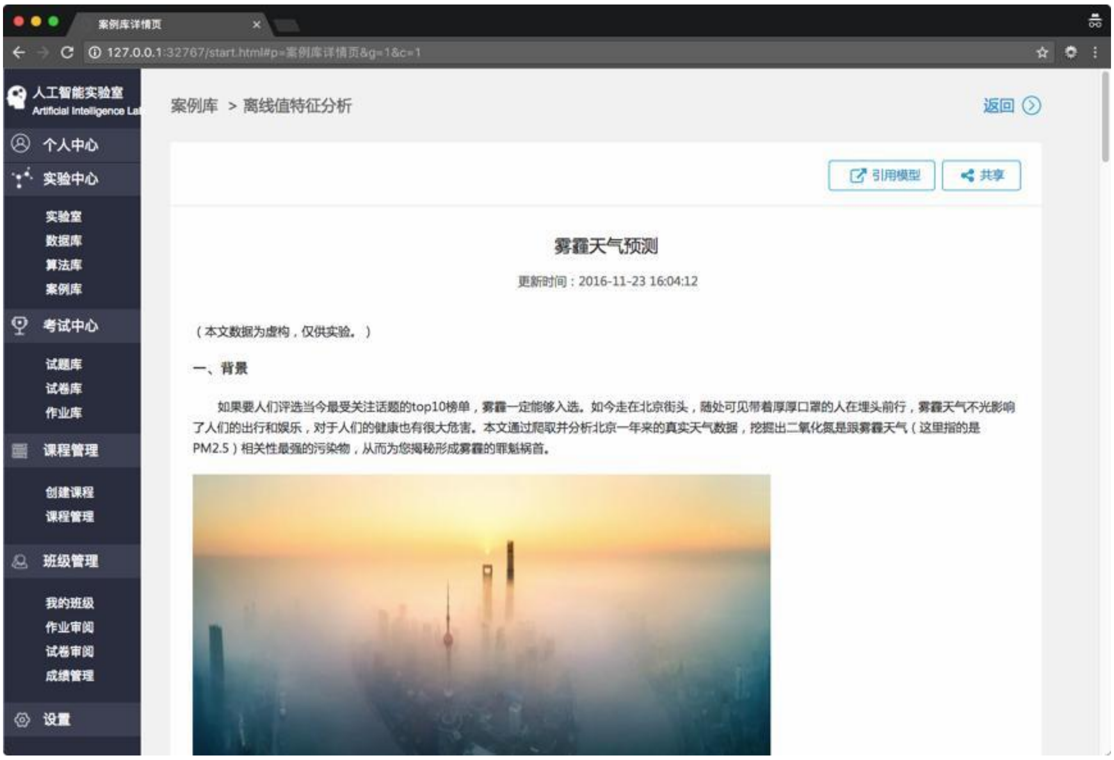
案例库列表可进入查看案例详细介绍；

可将案例中的算法模型引入至实验画布进行使用学习； 若自己上传的案例可共享给其他用户使用学习。

1) 案例库详细介绍

- 项目背景

介绍案例的开发背景以及需求，要达到的目标和效果。



（图 4-29）

■ 数据集介绍

项目使用的数据集要求和介绍，以及数据结构实例。

二、数据集介绍

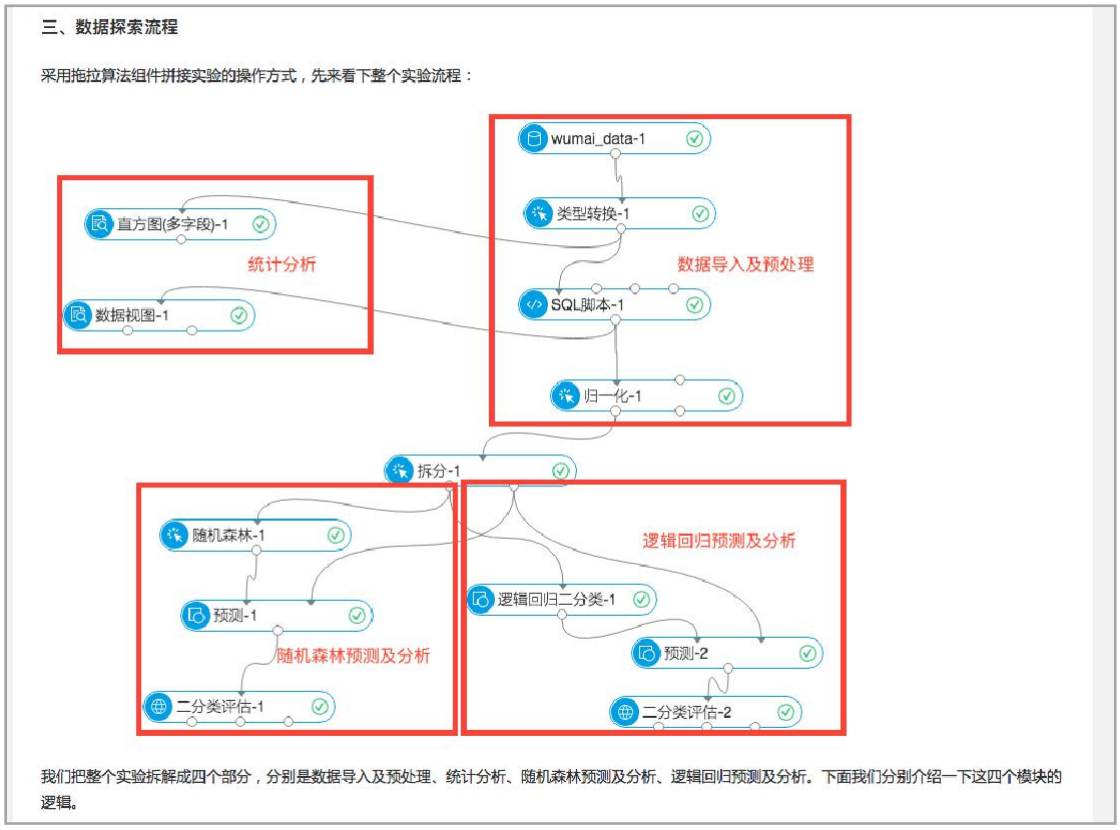
数据源：采集了2016全年的北京天气指标。
采集的是从2016年1月1号以来每个小时的空气指标，数据预览如下表：

省份	市场	销售额	省份	市场	销售额	省份	省份	市场	销售额
北京	北京	100	北京	北京	100	北京	北京	北京	100
北京	北京	100	北京	北京	100	北京	北京	北京	100
北京	北京	100	北京	北京	100	北京	北京	北京	100
北京	北京	100	北京	北京	100	北京	北京	北京	100
北京	北京	100	北京	北京	100	北京	北京	北京	100
北京	北京	100	北京	北京	100	北京	北京	北京	100
北京	北京	100	北京	北京	100	北京	北京	北京	100
北京	北京	100	北京	北京	100	北京	北京	北京	100

（图 4-30）

■ 算法界面操作

说明如何通过实验画布完成



（图 4-31）

案例算法流程设计。

- 数据加载及预处理

说明案例数据如何处理并与算法关联实例。

（图 4-32）

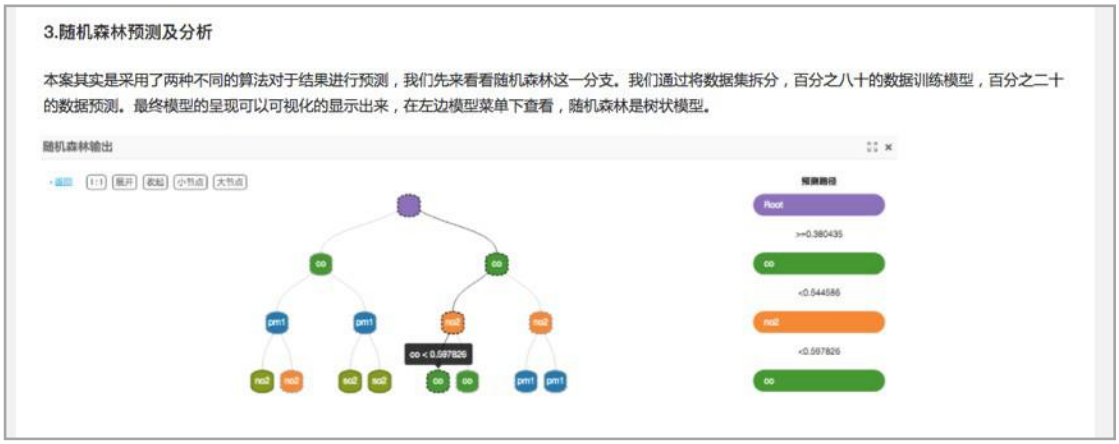
■ 统计分析

说明使用哪类可视化图形表示案例计算结果。

（图 4-33）

■ 案例核心思路

介绍案例使用了哪些算法，并详细介绍这些算法的核心思路以及相关代码



（图 4-34）

（图 4-35）

■ 结果分析和评估

展现案例最后达到的结算结果，以及如何评估计算正确。

四、结果分析和评估

因为经过归一化计算的逻辑回归算法有这样的特点，模型系数越大表示对于结果的影响越大，系数符号为正号表示正相关，负号表示负相关。我们看一下正号系数里pm10和no2最大。pm10和pm2只是颗粒尺寸大小不同，是一个包含关系，这里不考虑。剩下的no2(二氧化氮)对于pm2.5的影响最大。我们只要查阅一下相关文档，了解下哪些因素会造成no2的大量排放即可找出影响pm2.5的主要因素。下面网上是找到的关于no2排放的论述,文中说明了no2主要来自汽车尾气。

逻辑回归二分类

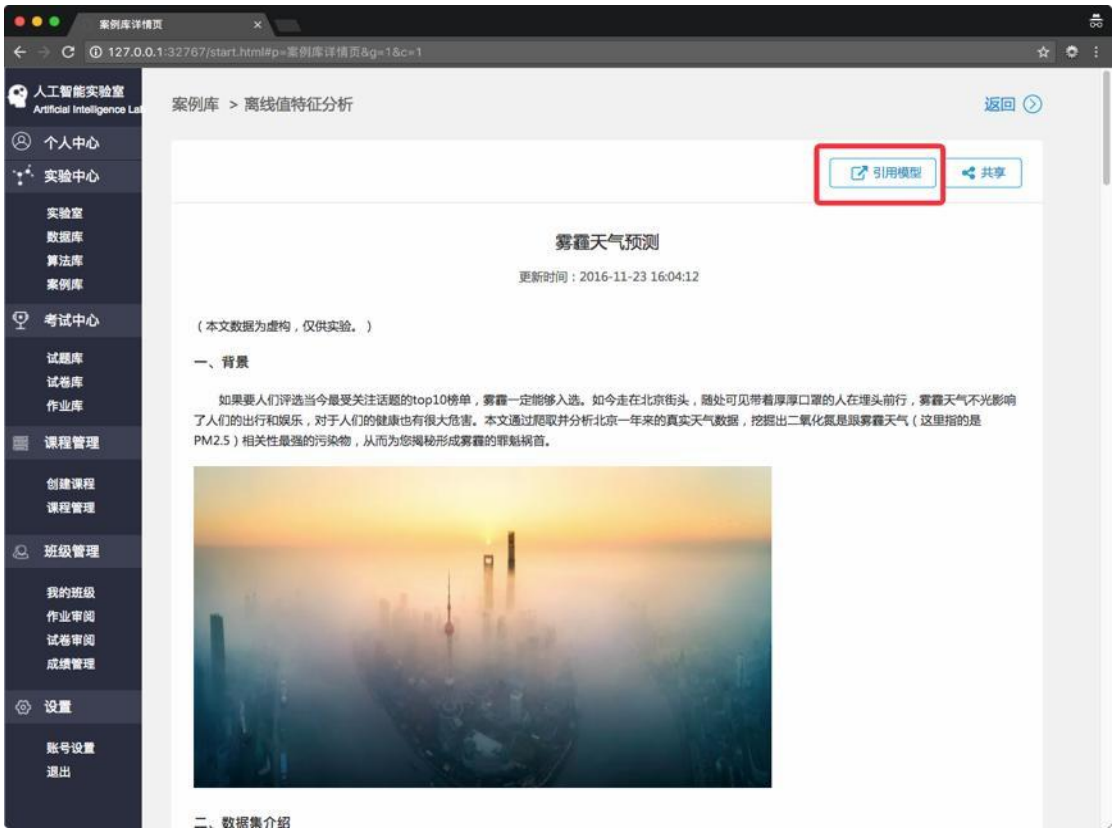
在输入数据为稀疏的时候，不显示 weight 全是 0 的特征

字段名	权重	
	1	0
pm10	18.32146628653672	-
so2	1.767062094833547	-
co	-0.2519492790928399	-
no2	10.95221282178011	-
常量	-16.66654139199668	0

（图 4-36）

3.3 引入案例算法

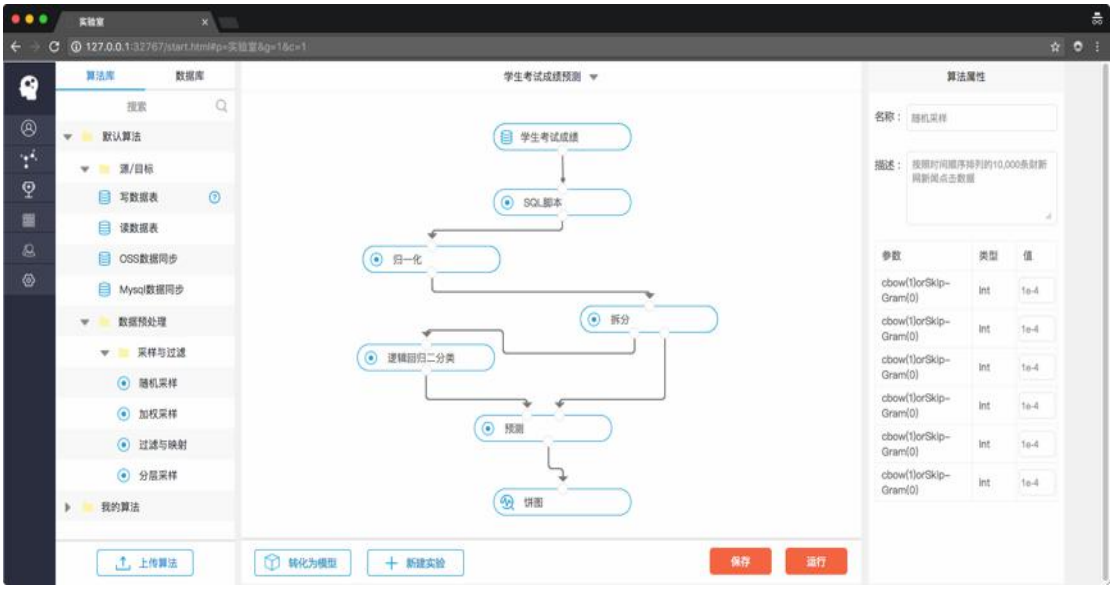
在案例列表页或案例详情页都可将案例中使用的算法设计模型引入至实验画布中。



（图 4-37）

■ 案例算法模型引入实验画布

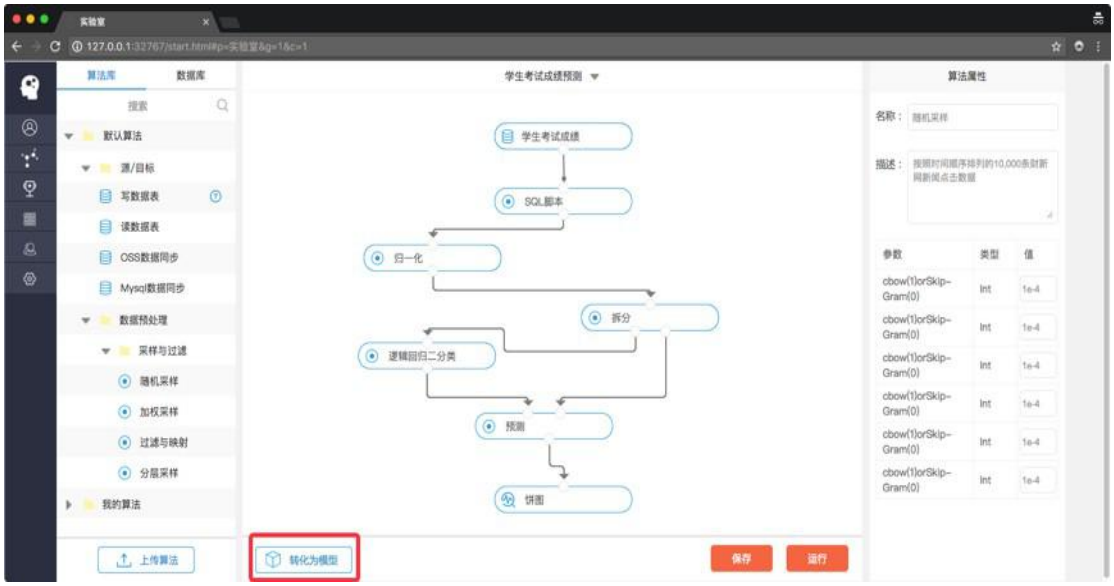
引入后，可对案例的算法流程进行操作，替换数据或修改相关配置，便于学习和研究。



（图 4-38）

将实验算法流程转化为案例

同时，也可以通过实验画布中的【转化为案例】按钮将实验中的流程转化为项目案例。



（图 4-39）

■ 案例内容补充

当点击【转化为案例】按钮后会打开案例内容补充界面，需针对该案例进行内容方面的补充说明，其中包括：案例名称、背景介绍、数据集介绍、界面操作说明、数据预处理过程、核心算法及介绍、统计分析介绍和结果分析和评估等内容，便于归档或共享其他用户使用。

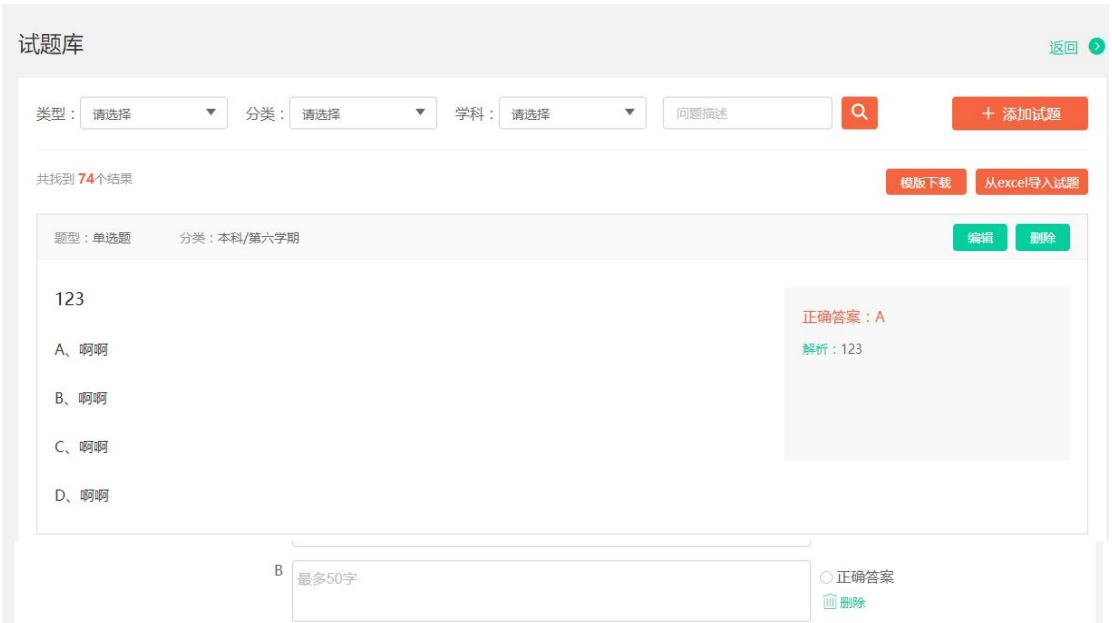
（图 4-40）

3.4 考试中心

1) 试题库

试题库，支持教师按课程分类、按题型（包含单选题、多选题和简答题）进行题目筛选，支持教师按关键字搜索相关题目，支持对已有的题目进行编辑、删除。支持教师创建新的试题，或下载模板批量上传试题至线上。创建新题时，支持设置课程分类、设置题型、编辑题目及选项信息，同时支持设置解析内容，学生提交试卷后支持查看解析内容，方便学生对知识点的理解。界面如下：

（图 4-41）



（图 4-42）

2) 试卷库/作业库

试卷库，支持教师按课程分类、考试形式（分为课程考试、统一考试和作业练习）对试卷进行筛选，支持教师按关键字搜索相关问卷。教师用户对不同发布状态试卷有不同使用权限，支持审阅已结束状态试卷，提供查看题目总量及分类数量、班级、总分设置、试卷时长设置等信息；未发布的问卷支持编辑、删除等操作，提供查看题目总量及分类数量、总分设置、试卷时长设置等信息；已发布的问卷支持查看、停用等操作，提供查看题目总量及分类数量、总分设置、试卷时长设置等信息。试卷库创建试卷时，支持设置试卷基本信息，如包含试卷名称、描述、答题时间、课程分类、考试形式、开始时间、班级设置等基本信息； 支持教师从试题库选择题目，也可直接创建新题；设置问卷分值时，支持自动平均题目分值； 支持教师实时总览各个题型的数量及题目总量；问卷支持按题目自动排序。界面如下：

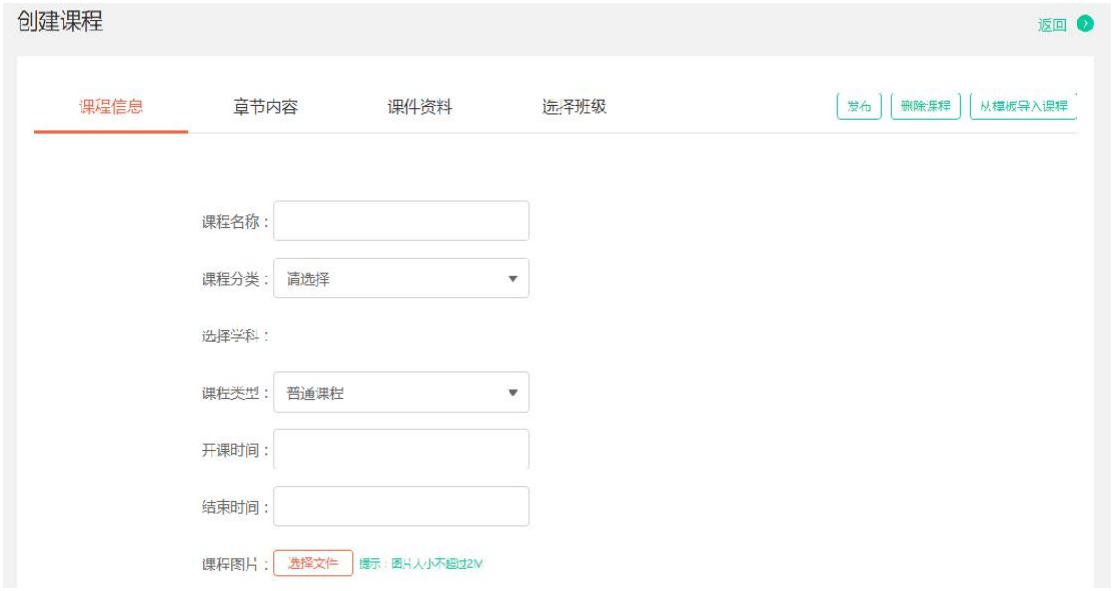
（图 4-43）

3.5 课程管理

教师用户管理课程，支持按分类、学科进行筛选，按课程名称进行搜索；快速定位到要管理的课程后，提供 7 项快速操作入口，包含预览、编辑、导入班级、查看班级、学员申请、成绩管理、实验报告等，方便管理；预览课程时，支持播放视频、快速进入实验等。界面如下：

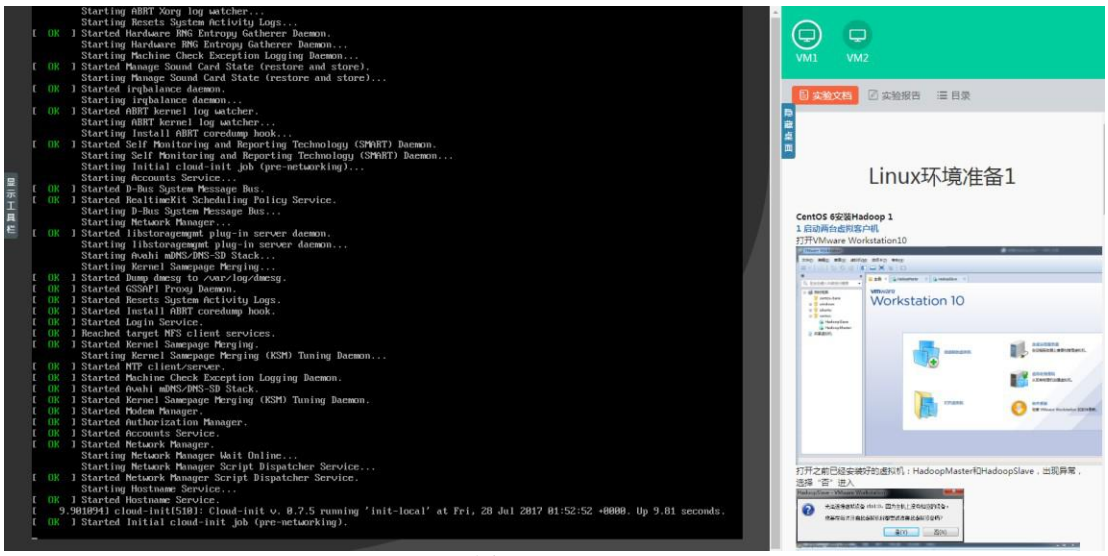
（图 4-44）

点击创建课程，进入创建课程界面，提供 4 项内容，包含课程信息、章节内容、课件资料和选择班级；提供 3 项操作，包含发布、删除、从模板导入课程。界面如下：



（图 4-45）

点击课程预览进入课程的所有视频和实验页面，在这里可以进行课程视频的观看和实验的操作。界面如下：

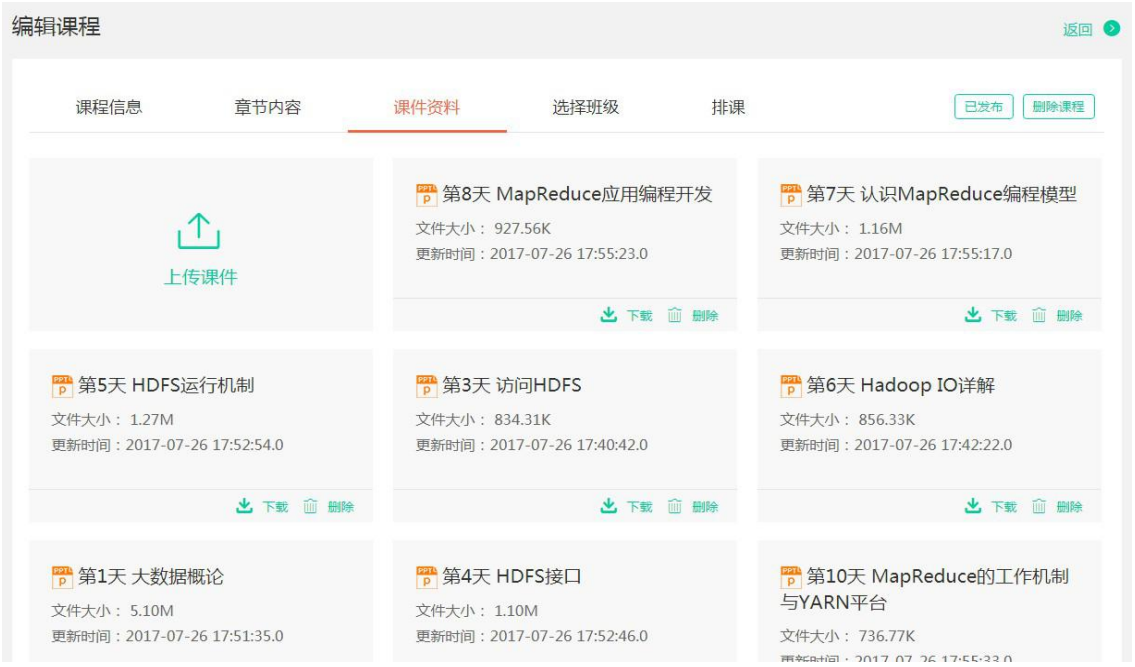


（图 4-46）

点击编辑可进入课程详细信息、章节的内容以及选择班级，并且发布的课程多了一项课程的安排，提供 2 种排课方式，包含手动排课和自动排课。界面如下：

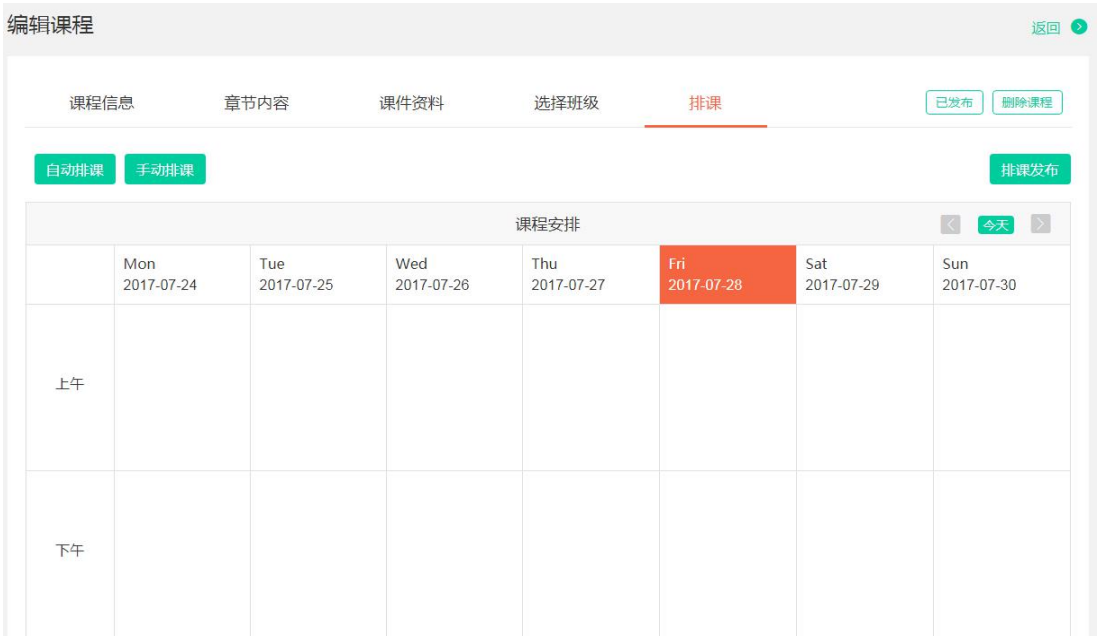
（图 4-47）

（图 4-48）



（图 4-49）

（图 4-50）

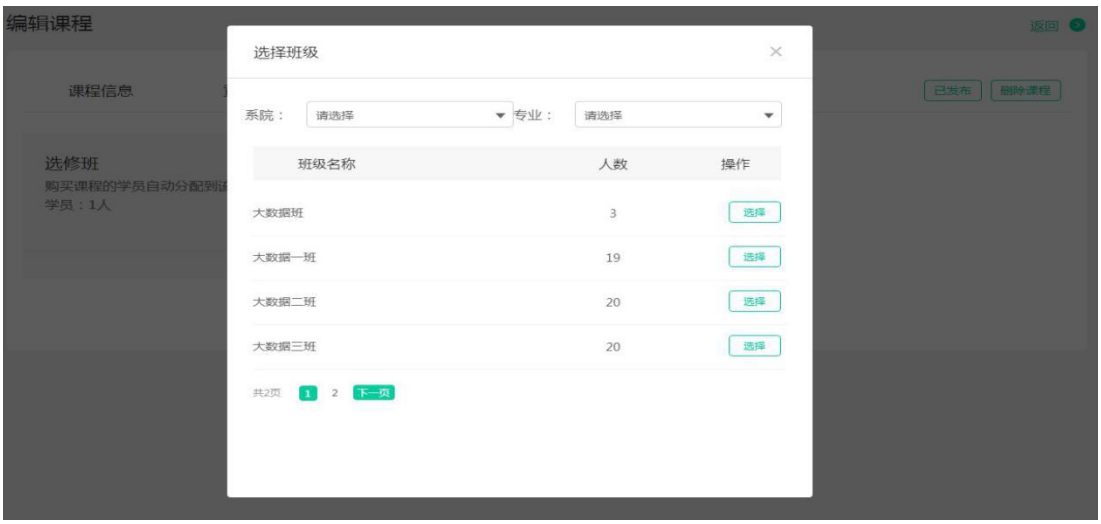


（图 4-51）

在管理课程中点击班级的导入则进入选择班级页面进行班级的选择和删除操作。界面如下：

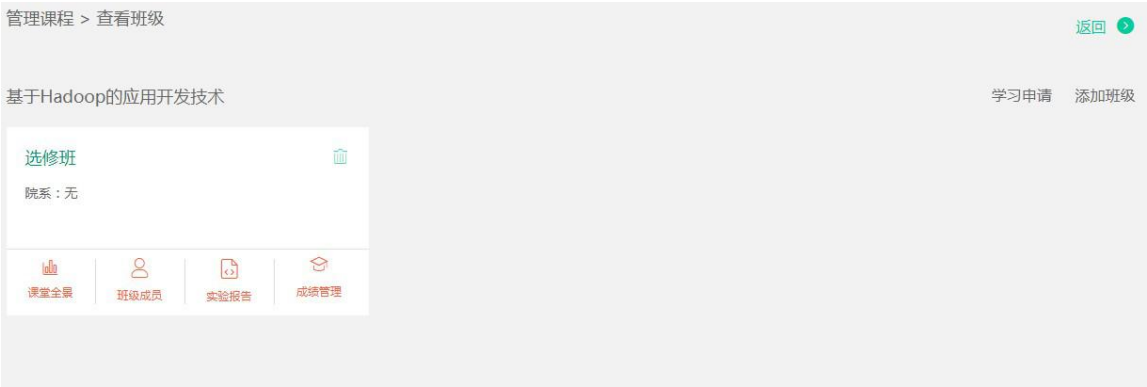
（图 4-52）

点击选择班级弹窗进行班级的选择班级选择时可以进行班级的院系和专业进行筛选。界面如下：



（图 4-53）

在管理课程中点击查看班级进入课程所关联的班级，界面如下：

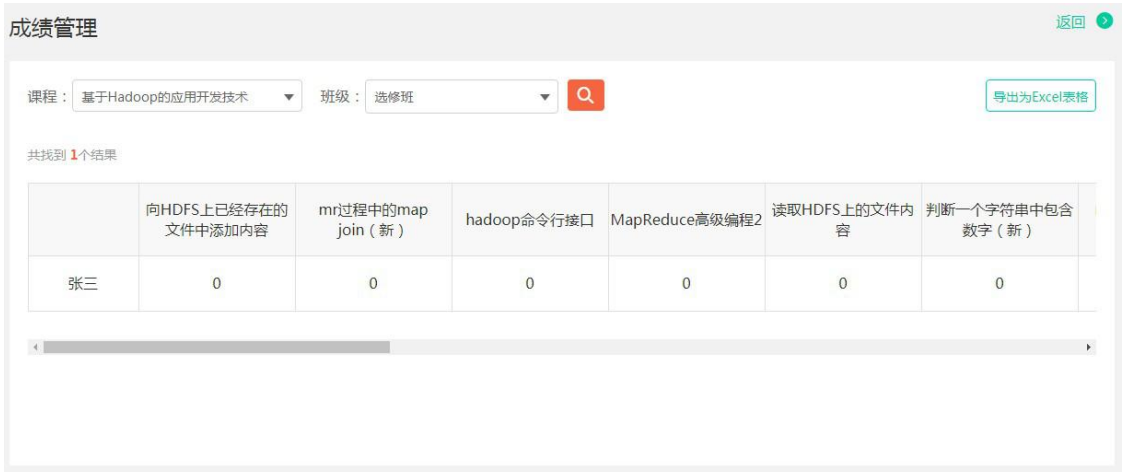


（图 4-54）

在管理课程中点击申请考核进入本课程下的所有学生申请的请求，可以通过或者拒绝。界面如下：

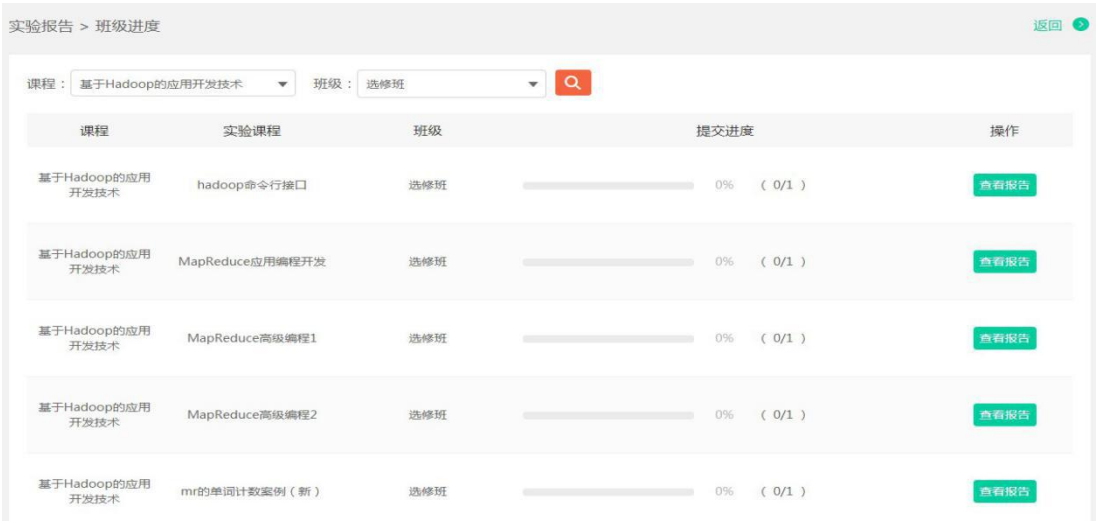
（图 4-55）

在课程管理中点击成绩管理是这个课程下所有学生的所做的实验成绩进行管理，界面如下：



（图 4-56）

在课程管理中点击实验报告进入这个课程下学生所做实验的报告的一个管理，界面如下：



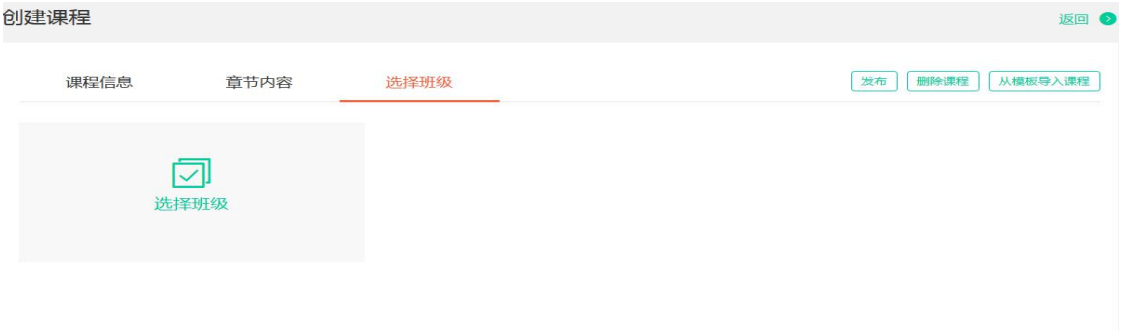
（图 4-57）

1) 创建课程

教师创建课程时，支持以向导的形式自主创建课程，同时支持从模板导入课程，快速创建。创建课程提供 4 项内容，包含课程信息、章节内容、课件资料和选择班级。界面如下：

（图 4-58）

当你输入完章节名称和章节内容后点击保存后会增加添加视频或者实验的界面，界面如下：

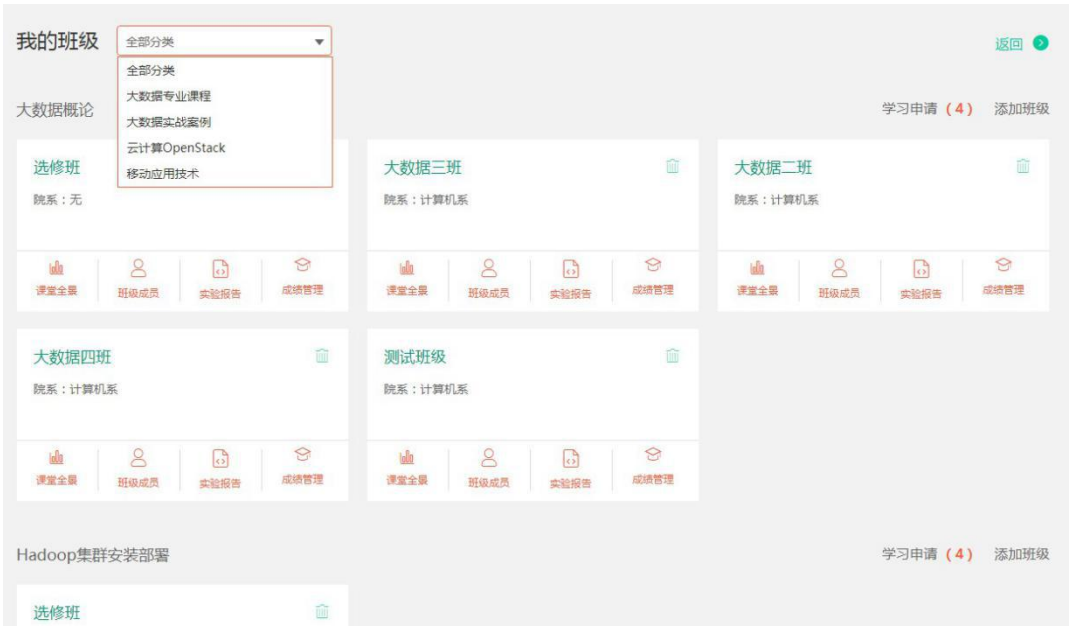


创建课程可以从模板课程中进行添加，界面如下：

3.6 班级管理

1) 我的班级

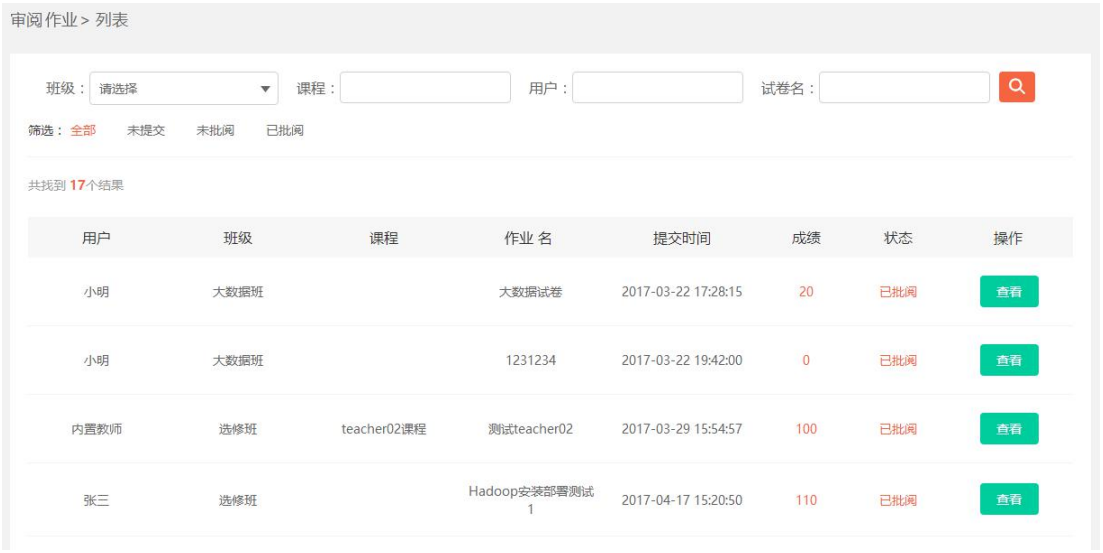
我的班级是自己所创建的课程下所有的班级，支持按照课程的类型进行课程的筛选，提供 4 项操作，包含删除班级、管理班级人员、班级的实验报告和成绩进行管理。界面如下：



（图 4-64）

2) 作业审阅/试卷审阅

教师审阅作业，支持按班级、课程、用户、试卷名、状态进行筛选，筛选结果包含 7 个字段信息，即用户名、班级、课程、试卷名、提交时间、成绩和状态。审阅试卷支持在线评分和在线批注。界面如下：



（图 4-65）

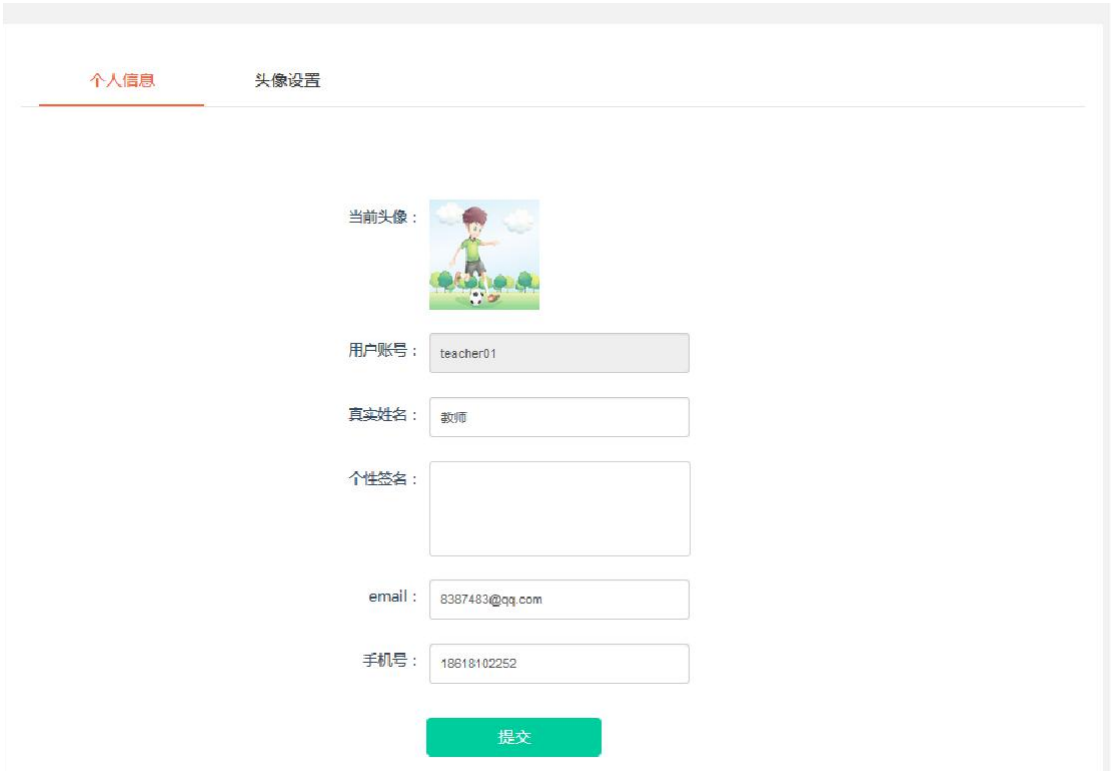
3) 成绩管理

成绩管理可以进入自己所有创建课程的所有同学所做实验的一个管理并且对他们的成绩进行操作。提供 4 项操作，包含筛选、搜索、修改成绩、导出为 Excel 格式。界面如下：

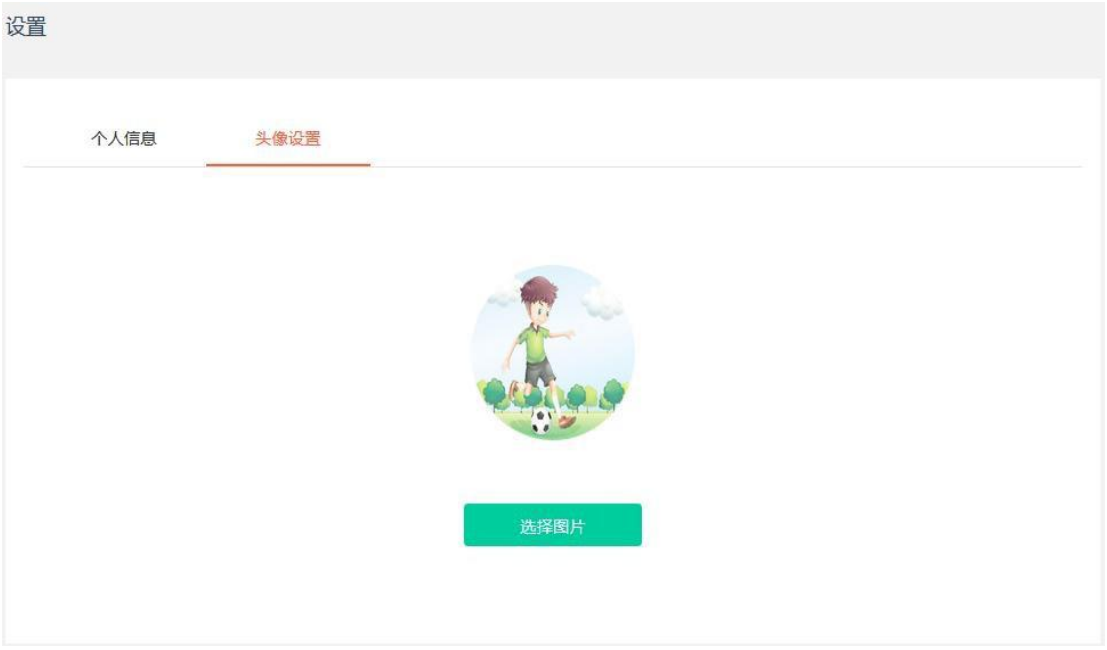
（图 4-66）

1.1 用户设置

设置是对个人信息的一个编辑，包含对个人信息和头像的编辑。


个人信息
头像设置
当前头像：
用户账号：
teacher01
真实姓名：
教师
个性签名：
email：
8387483@qq.com
手机号：
18618102252
提交

（图 4-67）


设置
个人信息
头像设置
选择图片

（图 4-68）

（四）人工智能训练包

人工智能教学与实验资源包是提供的若干个典型的人工智能技术实验资源，提供实验指导手册、实验数据源、实验过程中所需的大数据分析软件、实验参考示例代码和运行结果。学生可以在实验资源包的基础上完成仿真应用实验，完成创新创业实验和模拟大赛环境等相应的实验实训环节。

（1）基础算法训练集

普开人工智能实验室基础算法训练集主要是数据预处理算法、特征工程算法、统计分析算法、机器学习算法等等。

数据预处理（data preprocessing）是指在主要的处理以前对数据进行的一些处理。如对大部分地球物理面积性观测数据在进行转换或增强处理之前，首先将不规则分布的测网经过插值转换为规则网的处理，以利于计算机的运算。另外，对于一些剖面测量数据，如地震资料预处理有垂直叠加、重排、加道头、编辑、重新取样、多路编辑等。

特征工程，其本质是一项工程活动，目的是最大限度地从原始数据中提取特征以供算法和模型使用。

统计分析是指运用统计方法及与分析对象有关的知识，从定量与定性的结合上进行的研究活动。它是继统计设计、统计调查、统计整理之后的一项十分重要的工作，是在前几个阶段工作的基础上通过分析从而达到对研究对象更为深刻的认识。它又是在一定的选题下，集分析方案的设计、资料的搜集和整理而展开的研究活动。系统、完善的资料是统计分析的必要条件。

运用统计方法、定量与定性的结合是统计分析的重要特征。随着统计方法的普及，不仅统计工作者可以搞统计分析，各行各业的工作者都可以运用统计方法进行统计分析。只将统计工作者参与的分析活动称为统计分析的说法严格说来是不正确的。提供高质量、准确而又及时的统计数据和高层次、有一定深度、广度的统计分析报告是统计分析的产品。从一定意义上讲，提供高水平的统计分析报告是统计数据经过深加工的最终产品。

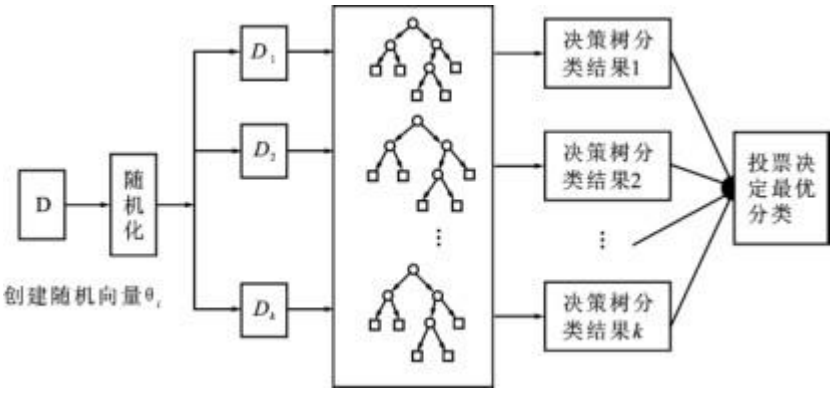
机器学习是近 20 多年兴起的一门多领域交叉学科，涉及概率论、统计学、逼近论、凸分析、算法复杂理论等多门学科。机器学习理论主要是设计和分析一些让计算机可以自动“学习”的算法。即从数据中自动分析获得规律，并利用规律对未知数据进行预测的算法。

机器学习是对能通过经验自动改进的计算机算法的研究。

1.1 随机森林模型

随机森林模型（Random Forest Model）是由许多决策树组合而成的分类模型，且参数集是独立同分布的随机向量，在给定自变量下，每个决策树分类模型都有一票投票权来选择最优的分类结果。

随机森林模型的基本思想是：首先，利用 bootstrap 抽样从原始训练集抽取k 个样本，且每个样本的样本容量都与原始训练集一样；其次，对 k 个样本分别建立 k 个决策树模型，得到 k 种分类结果；最后，根据 k 种分类结果对每个记录进行投票表决决定其最终分类，如下图所示。



通过构造不同的训练集增加分类模型间的差异，从而提高组合分类模型的外推预测能力。通过 k 轮训练，得到一个分类模型序列，再用它们构成一个多分类模型系统，该系统的最终分类结果采用简单多数投票法。

1.2 朴素贝叶斯

朴素贝叶斯模型（Naïve Bayesian Model，NBM）发源于古典数学理论，是有着坚实的数学基础及稳定分类效率的分类算法。朴素贝叶斯的基本思想是：首先，确定样本的特征属性及类型，并对训练样本进行属性划分，假定各个特征属性是条件独立的；然后，统计各类样本的发生概率，并通过贝叶斯定理计算各单属性下不同属性值的分类概率；对于给出的待分类项，求解在该项属性出现的条件下各类别发生的概率，比较各类别概率，将概率最高的类别作为此待分类项的类别。朴素贝叶斯模型参数较少，算法简单明了，建模速度快，但对缺失数据不太敏感。

1.3 决策树

决策树(Decision Tree）是在已知各种情况发生概率的基础上，通过构成决策树来求取净现值的期望值大于等于零的概率，评价项目风险，判断其可行性的决策分析方法，是直观运用概率分析的一种图解法。由于这种决策分支画成图形很像一棵树的枝干，故称决策树。在机器学习中，决策树是一个预测模型，他代表的是对象属性与对象值之间的一种映射关系。Entropy = 系统的凌乱程度，使用算法 ID3, C4.5 和 C5.0 生成树算法使用熵。这一度量是基于信息学理论中熵的概念。

决策树是一种树形结构，其中每个内部节点表示一个属性上的测试，每个分支代表一个测试输出，每个叶节点代表一种类别。

分类树（决策树）是一种十分常用的分类方法。他是一种监管学习，所谓监管学习就是给定一堆

样本，每个样本都有一组属性和一个类别，这些类别是事先确定的，那么通过学习得到一个分类器，这个分类器能够对新出现的对象给出正确的分类。这样的机器学习就被称之为监督学习。

（2）深度学习算法训练集

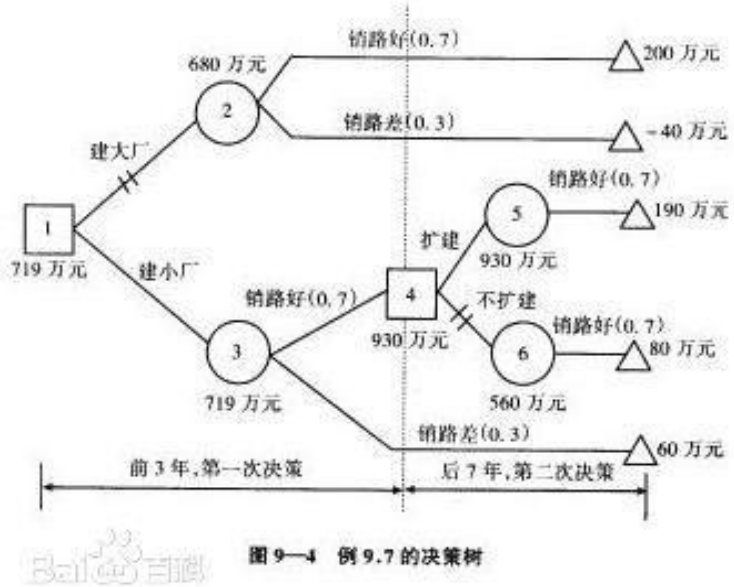


图 9—4 例 9.7 的决策树

普开人工智能实验室深度学习算法训练集主要是基于深度学习的相关算法训练集。

深度学习（Deep Learning）是人工神经网络的延展，透过模拟人类神经网络结构，使用包含复杂结构和由多重非线性变换构成的多个处理层对数据进行高层抽象等一系列算法，建立具有数个隐层的多层感知网络，实现各种模式的识别和认知，对图像识别、语音识别、自然语言处理、药物研发等领域的发展皆有突出贡献。

深度学习模型的基础单元为感知器，每个特征以不同的权重传递至下一层神经元，结合偏差组成新的、抽象化的特征。随后该特征将输入激活函数中进行非线性的转换，对提升模型的识别能力作用不大的会被忽略，反之成为下一层神经元的输入特征。透过不断地调整各特征在网络中的权重，进行表征学习，构成更具可分性的模式识别模型。

有别于大部分分类算法，基于可观测到的特征，拟合一个将观察数据连接的近似曲线， 深度学习模型更贴近于人类掌握知识和技能的方式。以图像识别为例，人类分辨图像展示的内容不只依靠一套标准的特征集（例如每个像素的 RGB 值、某种颜色所占的比例），还会借助更为抽象的观念，包括轮廓、主要组件等，计算机的认知过程也不应止步于前者。通过深度学习相关算法，这些较高层次的概念可从输入特征中推衍提炼出来，因此适合应用于样本量大、可观测的特征较小、特征分类能力较弱的数据集，对典型的分类问题也较其他分类算法有一定的性能提升。

深度学习的算法也是无监督或半监督式学习算法，用来处理存在少量未标识数据的大数据集。常见的深度学习算法包括：

- 受限波尔兹曼机（Restricted Boltzmann Machine）

- 深度信念网络（Deep Belief Networks）
- 卷积神经网络（Convolutional Neural Networks）
- 堆栈式自动编码器（Stacked Auto-encoders）

1.4 深度信念网络（Deep Belief Networks）

分别单独无监督地训练每一层RBM 网络，确保特征向量映射到不同特征空间时，都尽可能多地保留特征信息；

在 DBN 的最后一层设置 BP 网络，接收 RBM 的输出特征向量作为它的输入特征向量， 有监督地训练实体关系分类器。而且每一层 RBM 网络只能确保自身层内的权值对该层特征向量映射达到最优，并不是对整个 DBN 的特征向量映射达到最优，所以反向传播网络还将错误信息自顶向下传播至每一层 RBM，微调整个 DBN 网络。RBM 网络训练模型的过程可以看作对一个深层 BP 网络权值参数的初始化，使 DBN 克服了 BP 网络因随机初始化权值参数而容易陷入局部最优和训练时间长的缺点。

上述训练模型中第一步在深度学习的术语叫做预训练，第二步叫做微调。最上面有监督学习的那一层，根据具体的应用领域可以换成任何分类器模型，而不必是BP 网络。

1.5 卷积神经网络（Convolutional Neural Networks）

卷积神经网络是人工神经网络的一种，已成为当前语音分析和图像识别领域的研究热点。它的权值共享网络结构使之更类似于生物神经网络，降低了网络模型的复杂度，减少了权值的数量。该优点在网络的输入是多维图像时表现的更为明显，使图像可以直接作为网络的输入，避免了传统识别算法中复杂的特征提取和数据重建过程。

convolution 和 pooling 的优势为使网络结构中所需学习到的参数个数变得更少，并且学习到的特征具有一些不变性，比如说平移，旋转不变性。以 2 维图像提取为例，学习的参数个数变少是因为不需要用整张图片的像素来输入到网络，而只需学习其中一部分 patch。而不变的特性则是由于采用了 mean-pooling 或者 max-pooling 等方法。

CNN 是第一个真正成功训练多层网络结构的学习算法。它利用空间关系减少需要学习的参数数目以提高一般前向 BP 算法的训练性能。CNN 作为一个深度学习架构提出是为了最小化数据的预处理要求。在 CNN 中，图像的一小部分（局部感受区域）作为层级结构的最低层的输入，信息再依次传输到不同的层，每层通过一个数字滤波器去获得观测数据的最显著的特征。

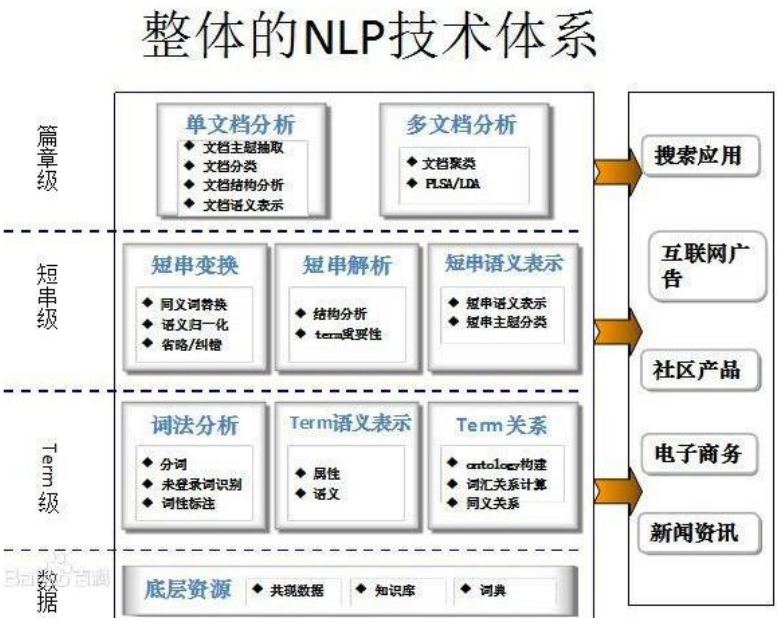
2.0 行业应用训练集

提供行业典型人工智能项目应用，项目涉及自然语言处理、机器视觉、人脸识别三大领域，每个项目案例都配备项目指导手册、行业数据、算法及项目代码。

1.6 自然语言处理

自然语言处理是计算机科学领域与人工智能领域中的一个重要方向。它研究能实现人与计算机之间用自然语言进行有效通信的各种理论和方法。自然语言处理是一门融语言学、计算机科学、数学于一体的科学。因此，这一领域的研究将涉及自然语言，即人们日常使用的语言，所以它与语言学的研究有着密切的联系，但又有重要的区别。自然语言处理并不是一般地研究自然语言，而在于研制能有效地实现自然语言通信的计算机系统，特别是其中的软件系统。因而它是计算机科学的一部分。

自然语言处理（NLP）是计算机科学，人工智能，语言学关注计算机和人类（自然）语言之间的相互作用的领域。



1.7 机器视觉

机器视觉就是用机器代替人眼来做测量和判断。机器视觉系统是指通过机器视觉产品（即图像摄取装置，分 CMOS 和 CCD 两种）将被摄取目标转换成图像信号，传送给专用的图像处理系统，根据像素分布和亮度、颜色等信息，转变成数字化信号；图像系统对这些信号进行各种运算来抽取目标的特征，进而根据判别的结果来控制现场的设备动作。

人工智能能使机器能够担任一些需要人工处理的工作。而这些工作需要做一定的决策， 要求机

器能够自行的根据当时的环境做出相对较好的决策。这就需要计算机不仅仅能够计算， 还能够拥有一定得智能。而要对周围的环境进做出好的决策就需要对周边的环境进行分析， 即要求机器能够看到周围的环境，并能够理解它们。就像人做的那样。所以机器视觉是人工 智能中非常重要的一个领域。

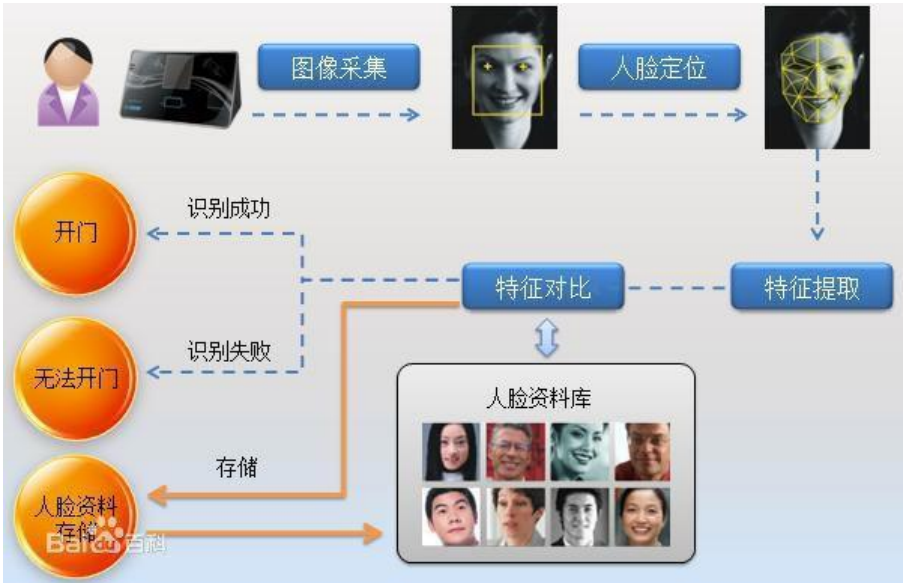
机器视觉在许多人类视觉无法感知的场合发挥重要作用，如精确定律感知、危险场景感知、不可见物体感知等，机器视觉更突出他的优越性。现在机器视觉已在一些领域的到应用， 如零件识别与定位，产品的检验，移动机器人导航遥感图像分析，安全减半、监视与跟踪， 国防系统等。它们的应用于机器视觉的发展起着相互促进的作用。

1.8 人脸识别

人脸识别特指利用分析比较人脸视觉特征信息进行身份鉴别的计算机技术。人脸识别是一项热门的计算机技术研究领域，人脸追踪侦测，自动调整影像放大，夜间红外侦测，自动调整曝光强度；它属于生物特征识别技术，是对生物体（一般特指人）本身的生物特征来区分生物体个体。

人脸识别技术是基于人的脸部特征，对输入的人脸图象或者视频流。首先判断其是否存在人脸，如果存在人脸，则进一步的给出每个脸的位置、大小和各个主要面部器官的位置信息。并依据这些信息，进一步提取每个人脸中所蕴涵的身份特征，并将其与已知的人脸进行对比，从而识别每个人脸的身份。

更深层次的是和大数据打通。尤其人脸大数据，无论在日常生活，还是商业运作上都是语音、动作之后最重要的数据之一，它更能够将个人大数据实现更大化的整合，甚至重建信用体系规则。



1.9 语音识别

语音识别是一门交叉学科。近二十年来，语音识别技术取得显著进步，开始从实验室走向市场。人们预计，未来 10 年内，语音识别技术将进入工业、家电、通信、汽车电子、医疗、家庭服务、消费电子产品等各个领域。语音识别听写机在一些领域的应用被美国新闻界评为 1997 年计算机发展十件大事之一。很多专家都认为语音识别技术是 2000 年至 2010 年间信息技术领域十大重要的科技发展技术之一。语音识别技术所涉及的领域包括：信号处理、模式识别、概率论和信息论、发声机理和听觉机理、人工智能等等。

（五）物理层环境建设

（1）硬件环境

为满足大数据实验实训需求，以及本科生、研究生创新实验、创新创业大赛、互联网+ 大赛等比赛需求，必须新增基于业界最先进的硬件平台，采用企业级融合架构，针对不同软件单元的特性，对计算单元、网络单元和存储单元进行多层次重组和整合优化，具备高效融合、企业级 PAAS、安全稳定、性能卓越、敏捷易用等特点。包括计算、管理、交换等节点设备。

产品名称	配置说明
GPU 计算节点	配置 Intel Xeon 6126 处理器(2.6GHz/12c)/2666MHz/10.4GT*2 配置 32G ECC Registered DDR4 2666 内存*12 配置 300GB 2.5″ 15Krpm SAS 硬盘*2 配置 RAID 卡_L_8R0_9361-8i_1GB_HDM12G_PCIE3.0 配置网卡_INSPUR_5280M5_1543_1G_RJ_DC_4_XR_PH 配 置显卡_NV_16G_TESLA-P100_4096b_P_CAC*4 配置单端口 EDR 100Gbps HCA 卡配 置 1600W 1+1 冗余服务器电源

计算节点	配置 Intel Xeon E5-2620v3 处理器*2 配置 16GB DDR4 内存*10 配置 300GB SAS 热插拔硬盘*2，4TB SATA 硬盘*3 配置高性能 RAID 卡，支持 RAID0/1/5/10 配置千兆网口*2，万兆网口*2
管理节点	配置 Intel Xeon E5-2603v3 处理器*2 配置 16GB DDR4 内存*4 配置 300GB SAS 热插拔硬盘*2，2TB SATA 硬盘*3 配置高性能 RAID 卡，支持 RAID0/1/5/10 配置千兆网口*2，万兆网口*2
交换节点	配置 48 个 10/100/1000Base-T 以太网端口 支持堆叠
万兆交换节点	配置 24 个万兆 SPF+接口配 置万兆光纤模块 支持堆叠
机柜	42U 标准服务器机柜 配置 2 个 8 口 PDU
KVM	配置 16 端口 显示屏：17 英寸 LED

（2）人工智能实验软件环境系统

服务器	计算节点：配置Intel Xeon E5-2620v3处理器*2；配置192GB DDR4内存；配置300GB SAS热插拔硬盘*2，4TB SATA硬盘*3；配置高性能RAID卡，支持RAID0/1/5/10；配置千兆网口*2
服务器	管理节点:配置Intel Xeon E5-2603v3处理器*2；配置64GB DDR4内存；配置300GB SAS热插拔硬盘*2，2TB SATA硬盘*3；配置高性能RAID卡，支持RAID0/1/5/10；配置千兆网口*4
GPU服务器	处理器：INTEL E5-2620v3系列处理器，最大支持2颗，最大支持160W TDP。 芯片组：高效C612芯片 内　存：最大支持2T(16 DIMM Slots)，频率最高支持2400MHz 硬　盘：最大支持8个3.5″SATA热插拔硬盘 阵　列：支持RAID 0、1、10（Only for Windows），可选独立阵列卡 网络I/O：集成2个(i350) 10/100/1000M自适应以太网口，一个专用远程管理口 GPU：最大支持4片NVIDIA Tesla P100/P40/K80/K40 光　驱：选配内置SATA接口厚DVD，可选外置USB光驱 电　源：2000W（1+1）高效节能服务器冗余电源 2 * CPU Intel E5-2620 V3（2.4G/15M/6C/12T/85W） 4 * 内存 32G DDR4 RECC 2133 2 * 硬盘 300G SAS 企业级 2.5 15000 3 * 硬盘 2T SATA 企业级 3.5 7200 2 * Tesla P40 24G 显存 1 * 阵列卡 SAS2208 6Gbps 1GB
交换机	配置24个10/100/1000Base-T以太网端口；支持堆叠
机柜	42U标准服务器机柜；含PDU

数据交换模块	1、系统提供Web服务访问，支持用户通过Web进行数据挖掘分析及HDFS上的数据查看、管理。 2、系统可支持多类型数据源的接入，同时提供有HDFS与数据库及用户本地系统的双向数据导入导出
数据预处理	1、支持如下的技术操作：过滤类型检查及去重、外键约束主键约束、缺值处理、空值域约束、去重、断行处理、去极值、caseWhen、计数区间化、数值区间化、字段类型转换、归一化、逆归一化、添加ID、属性交换、PCA、因子分析、最优离散化、Delete、Where、Select、Sort、JoinOnSpark、统计、计算生成列、计算幂次、groupby、列求和、基于主键的集合差、集合差、集合交并、分层抽样、随机抽样、系统抽样、整群抽样
数据挖掘模块	1、系统可支持处理结构化和非结构化的数据，支持针对特定业务数据类型补充处理逻辑。 2、支持多个计算框架同时运行，包括：MapReduce、Spark、Sqoop、Storm等。系统提供多类别分布式算法：包含有并行数据预处理ETL算法、统计分析、文本处理及并行化经典数据挖掘算法，支持用户针对实际业务逻辑所需，添加补充用户自行采取python、scala等编写的算法模型。 3、系统提供良好的可扩展性：系统以组件形式封装算法模型以提供用户使用，集合各类模型构建出可扩展的算法库。支持依据业务需要，在不影响系统运行的同时，动态添加业务相关算法以及其它功能逻辑满足后续任务需求。
算法模块	1、分类算法：svm、线性回归、逻辑回归、TextNaiveBayes、NaiveBayes、c45、CART、特征选择、CHAID、BP 2、聚类算法：K均值、LDA、DBScan、Birch、Clara 3、关联规则：Apriori、FPGrowth、时序关联规则、ITRule 4、数据探索：常用统计量、单变量分析、多变量分析、假设检验、自助法、主成分分析、statistics输出
操作分析模块	1、分析操作界面友好，支持用户通过可视化形式，结合操作画布，采取拖拽形式对组件进行操作，自定义数据挖掘分析任务。采取工作流图形式记录整个数据分析任务，以数据流图中数据流向表示数据挖掘分析任务流程方向，增强用户对数据分析挖掘的过程理解，实现实际数据挖掘分析任务。
任务实时监控模块	1、系统提供有两种任务监控界面：流程实时监控、Spark程序实时监控，支持用户对任务运行状态及数据流程分析进度进行实时查看。
案例模型模块	1、系统提供有工作流机制，支持以工作流图形式将数据挖掘任务存储，提供自身及其它人后续复用。支持用户将已经存储的工作流，读取至操作画布展现，包含所有相应组件及组件配置。支持用户对所拥有修改权限
权限管理模块	1、系统提供有调度流功能，支持用户自定义定制针对所需工作流的调度流程。 2、支持定制权限管理，包含数据资源使用管理权限、工作流程使用修改权限等
数据挖掘应用案例	1、关键词提取：系统导入信息为缺陷部分描述数据，为文本描述信息。系统平台通过调用文本分词、LDA等组件，依照逻辑依次连接，采用HMM、TF-IDF等方法从非结构化文本信息中，最终提取出结构化的关键词。 2、EMS量测数据预测：系统导入数据为EMS监控数据，包含电网中线路所检测到的所有信息。系统平台通过调用Select、GroupBy、SVM等组件，按照数据变换逻辑依次连接，以提取出有影响的因素，并划分为测试及预测两部分数据。而后根

	据数据建立模型，并最终完成对有功功率P的数据预测。 3、输变电设备状态聚类：系统导入数据为设备状态信息，为文本描述信息。系统平台通过调用数值化、birch等组件，按照分析逻辑依次连接，形成聚类分析流程，最终得到各类设备状态信息之间聚集关系。 4、油色谱主成分分析：系统导入数据为变压器中各时间段油色谱中各气体数据。系统平台通过调用统计、DC-PCA等组件，依照分析逻辑依次连接，基于分治法，采用分离合并策略将原有数据文件转换成数据矩阵并分解，从而最终得到权值最重的气体成分。 5、运营商解决方案预测：系统导入数据为甘肃省运营商受理业务处理数据。系统平台通过调用统计、BP等组件，依照分析逻辑依次连接，基于原有受理业务处理数据，建立模型，形成解决方案与其它相关属性的相应关系，最终能在得到其它相应属性值情况下预测所应该使用的处理解决方案。 6、农业商品价格预测：系统导入数据为各地各类农产品数据，其中包含位置、价格、种类等信息。系统平台通过调用数值化、回归建模等组件，依照分析逻辑依次连接，将原有非数值数据转换为数据值量化数据，从而建立回归模型，最终通过其它非价格属性得到商品价格。