**数据挖掘实践教学平台研究与实现**

**详细设计说明书**

目录

[1. 引言 3](#_Toc492195323)

[1.1 背景 3](#_Toc492195324)

[1.2 目的 3](#_Toc492195325)

[1.3 数据挖掘概览 3](#_Toc492195326)

[1.3.1 概念 3](#_Toc492195327)

[1.3.2 步骤 3](#_Toc492195328)

[1.3.3 系统需求 5](#_Toc492195329)

[2. 功能需求 5](#_Toc492195330)

[2.1 数据集管理部分 5](#_Toc492195331)

[2.2 学生信息管理部分 6](#_Toc492195332)

[2.3 实践任务管理部分 6](#_Toc492195333)

[2.4 数据挖掘算法库管理部分 7](#_Toc492195334)

[3. 系统总体设计 7](#_Toc492195335)

[3.1 构架 7](#_Toc492195336)

[3.1 系统结构 8](#_Toc492195337)

[3.2 系统流程 9](#_Toc492195338)

[3.2.1 登录流程 9](#_Toc492195339)

[3.2.2 系统功能选择流程 10](#_Toc492195340)

[3.2.3 数据集管理模块流程 12](#_Toc492195341)

[3.2.4 算法库管理流程 13](#_Toc492195342)

[3.2.5 学生管理流程 14](#_Toc492195343)

[3.2.6 实践任务管理流程 15](#_Toc492195344)

[4. 数据库设计与构架 16](#_Toc492195345)

[4.1 E-R图 16](#_Toc492195346)

[4.1.1 登录模块 17](#_Toc492195347)

[4.1.2 数据集管理 18](#_Toc492195348)

[4.1.3 算法库管理 20](#_Toc492195349)

[4.1.4 学生管理 21](#_Toc492195350)

[4.1.5 实践任务分配 22](#_Toc492195351)

[4.2 数据库关系模型设计 23](#_Toc492195352)

[4.3 数据库逻辑结构设计 23](#_Toc492195353)

[5. 模块设计 24](#_Toc492195354)

[5.1 详细模块设计 24](#_Toc492195355)

[5.1.1 登录注册模块 24](#_Toc492195356)

[5.1.2 数据集管理模块 25](#_Toc492195357)

[5.1.3 学生信息管理模块 27](#_Toc492195358)

[5.1.4 实践任务管理模块 28](#_Toc492195359)

[5.1.5 数据挖掘算法库管理部分 29](#_Toc492195360)

# 引言

## 背景

近年来，数据挖掘引起了信息产业界的极大关注，每时每刻产生的海量数据存在极具价值的分析发掘潜能。其目的是发现未知的关系和以数据拥有者可以理解并对其有价值的新颖方式来总结数据，广泛用于包括商务管理，生产控制，市场分析，工程设计和科学探索等各种应用。

## 目的

本系统意在提供一个为老师、学生提供一个实践教学平台，管理员可以在平台给学生分配指定类型的数据发掘任务，并同步追踪学生的数据发掘进度，进行有效的教学评估。学生可以根据任务在平台上传数据集，进行数据挖掘流程的分类、估计、预测、分析、总结工作，并上传期望进度，汇报发掘结果。系统会提供诸如图形、表格的数据呈现手段，帮助教学任务的较好完成。

## 数据挖掘概览

### 概念

数据挖掘是从大量的、不完全的、有噪声的、模糊的、随机的实际应用数据中，提取隐含在其中的、人们事先不知道的、但又是潜在有用的信息和知识的过程。数据挖掘可以描述为:按企业既定业务目标，对大量的企业数据进行探索和分析，揭示隐藏的、未知的或验证己知的规律性，并进一步将其模型化的有效方法。数据挖掘相近的同义词包括：数据融合、数据分析和决策支持等。

### 步骤

**1.业务理解：**

#### 确定业务目标：分析项目的背景，从业务视点分析项目的目标和需求，确定业务角度的成功标准；

#### 项目可行性分析：分析拥有的资源，条件和限制，风险估计，成本和效益估计；

#### 确定数据挖掘目标：明确确定数据挖掘的目标和成功标准，数据挖掘的目标和业务目标是不一样的，前者指技术上的，例如生成一棵决策树等；

#### 提出项目计划：对整个项目做一个计划，初步估计用到的工具和技术。

**2.数据理解：**

#### 收集原始数据：收集本项目所涉及到的数据，如有必要，把数据装入数据处理工具，并作一些初步的数据集成的工作，生成相应报告;

#### 描述数据：对数据做一些大致的描述，例如记录数、属性数等，给出相应报告；

#### 探索数据：对数据做简单的统计分析，例如关键属性的分布等；

#### 检查数据质量：包括数据是否完整、数据是否有错、是否有缺失值等问题。

**3.数据准备：**

数据清洗：

#### 数据的完整性：例如PM2.5的属性缺少风向、气压等属性；

#### 数据的唯一性：例如不同来源的数据出现重复的情况；

#### 数据的权威性：例如同一个指标下出现多个不同来源的数据且数值不同；

#### 数据的合法性：例如获取的数据与常识不符，如风向出现南北的方向；

#### 数据的一致性：例如不同来源的不同指标，实际内涵是一样的，如气压和压强,或者同一指标内容不一致；

针对方案：

#### 通过其他数据补全：如通过身份证号补全一个人的籍贯，出生年月；

#### 通过前后数据补全：如通过前后时间均值来补全缺失时间；

#### 舍弃补不全的数据：从列表剔除残缺的数据，但仓库里予以保留；

数据的唯一性：

#### 按主键去整理：保留唯一数据；

#### 按算法去整理：对不同来源的数据按一定规则合并或者只保留一个；

#### 数据的权威性：对数据来源进行权威性分级；

#### 数据的合法性：

#### 值域合法性：设定数值上下界来确保数据值的合法性；

#### 值类型合法性：如日期格式必须为2017/5/6；

（2）数据集成：

在原有数据的基础上生成新的属性或记录，利用表连接等方式将几个数据集合并在一起，把数据转换成适合数据挖掘处理的格式，将数据放到数据仓库中。

**4.建立模型阶段：**

#### 数据挖掘算法和参数选取：确定数据挖掘算法和参数，可能会利用多个算法；

#### 测试方案设计：设计某种测试模型的质量和有效性的机制；

#### 模型训练：在准备好的数据集上运行数据挖掘算法，得出一个或者多个模型；

#### 模测试评估：根据测试方案进行测试，从数据挖掘技术的角度确定数据挖掘目标是否成功。

**5.模型评估：**

#### 结果评估：从商业角度评估得到的模型，甚至实际试用该模型测试其效果；

#### 过程回顾：回顾项目的所有流程，确定每一个阶段都没有失误；

#### 确定下一步工作：根据结果评估和过程回顾得出的结论，确定是部署该挖掘模型还是从某个阶段开始重新开始。

### 系统需求

# 功能需求

本系统需要为教师提供一个友好的数据挖掘教学平台，让学生可以体验到数据挖掘基本流程。系统应包括数据集管理功能，学生基本信息概览，实践任务分配以及数据挖掘算法纳入等几个大方向的功能。学生可以在系统进行下载数据集，查看所分配的任务，进行数据挖掘任务几个过程。管理员可以管理学生信息，进行任务分配，上传、修改数据集等。

具体的功能包括：

#### 首页登录部分

#### 注册用户信息：学生可以通过页面注册自己的账号及密码。

#### 用户登录：系统根据账号及密码进行登录授权

#### 展示系统基本信息：包括系统的介绍、功能、联系方式、分享方式等。

## 数据集管理部分

学生：

#### 浏览数据集库：可以查看所展示的所有数据集的种类、数量、主题等概括性基本信息，支持根据不同的信息进行分类查看。

#### 浏览数据集：浏览数据集基本描述信息及具体内容，基本的描述信息包括：数据集的名称、年份、实例数、所属的类型（分类、聚类）等。数据集的具体内容可以查询，并且能够按照不同的条目（如年份、类型）进行分类排序浏览。

管理员：

#### 上传数据集：上传数据集的文件，以文件或二进制流的方式保存在服务器或mysql数据库中去。

#### 下载数据集：提供数据集的下载功能。

#### 删除数据集：删除已上传的数据集。提示相应的删除提示：如已被关联的数据集删除警告、数据量大的评估警告。

#### 更新数据集：可更新数据集的基本信息、属性、批量删除部分实例。

#### 文件读入数据集信息：根据管理员上传的属性描述文件，读入数据集的基本信息。

#### 浏览数据集：可以查看所展示的所有数据集的概括性基本信息，支持根据不同的信息进行分类查看。

## 学生信息管理部分

管理员：

#### 添加学生信息：管理员对学生信息进行添加，具体内容包括学生的姓名、学号、班号、专业等。

#### 修改学生信息：管理员对学生信息进行修改。

#### 删除学生信息：管理员对学生信息进行删除。

#### 查询学生信息：管理员对学生信息进行查询。

## 实践任务管理部分

管理员：

#### 任务数据集指定：合理默认为几个任务指定数据集库里的数据集。

#### 任务学生分组指定：合理为默认任务指定几组不同的学生，每组学生每次只能分配一组任务，一个任务可以由多组学生参与。

#### 任务算法配置：为不同类型任务配置适量的算法。

#### 新增任务：新增数据挖掘任务。

#### 修改任务：对任务所分配的数据集、学生、算法进行添加、删除、修改。

#### 跟踪进度：记录学生执行数据挖掘任务的进度，呈现学生任务的阶段。

#### 挖掘评测：调用学生完成任务后得到的结果存储记录，比较结果算法、算法参数、实验方法等因素造成的差异。

学生：

#### 查看分配任务：学生查看系统分配给自己的任务。

#### 提交任务：对阶段性的任务进行提交、总结。

#### 查看提交结果：对已提交的结果进行查看。

## 数据挖掘算法库管理部分

#### 算法浏览：展示系统内数据挖掘算法。支持分类、聚类算法分类浏览，

#### 算法呈现：对具法算法进行阅读。包括算法参数、类型、接口的综合描述。

# 系统总体设计

## 构架

## 系统结构



图 3.1系统功能模块图

## 系统流程

### 登录流程



图 3.2 登录流程图

### 系统功能选择流程



图 3.3 系统功能流程图

### 数据集管理模块流程

 图 3.4 数据集管理模块流程图

### 算法库管理流程

图 3.5 算法库管理模块流程图

### 学生管理流程



图 3.6学生管理模块流程图

### 实践任务管理流程



图 3.7 实践任务管理模块流程图

# 数据库设计与构架

## E-R图

本模块主要展示数据库核心的ER设计关系（限于篇幅，或有细节关系没列出）

### 登录模块



图 4.1权限验证核心ER图

### 数据集管理



图 4.2 数据集管理核心ER图

### 算法库管理

图 4.3 算法库核心ER图

### 学生管理



图 4.4 学生管理核心ER图

### 实践任务分配

图 4.5 数据挖掘实践任务分配核心ER图

## 数据库关系模型设计

## 数据库逻辑结构设计

### 权限

1、teacher

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 字段名称 | 数据类型 | 长度 | 允许为空 | 中文说明 | 备注 |
| TeacherId | int | 11 | 否 | 学号 | 主键 |
| TeacherName | Varchar | 20 | 否 | 教师姓名 |  |

2、student

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 字段名称 | 数据类型 | 长度 | 允许为空 | 中文说明 | 备注 |
| studentId | int | 11 | 否 | 学号 | 主键 |
| classId | int | 11 | 否 | 班号 |  |
| professtions | Varchar | 20 | 否 | 专业 |  |

3、user

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 字段名称 | 数据类型 | 长度 | 允许为空 | 中文说明 | 备注 |
| userId | Varchar | 32 | 否 | 用户编号 | 主键 |
| userName | Varchar | 20 | 否 | 用户姓名 |  |
| password | Varchar | 45 | 否 | 密码 |  |
| userAcessStatus | Varchar | 45 | 否 | 用户权限状态 |  |

4、role

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 字段名称 | 数据类型 | 长度 | 允许为空 | 中文说明 | 备注 |
| roleId | Varchar | 32 | 否 | 学号 | 主键 |
| roleName | Varchar | 20 | 否 | 班号 |  |
| description | Varchar | 20 | 否 | 专业 |  |
| isAvailable | boolean |  | 是 | 可否为空 | 默认可以 |
|  |  |  |  |  |  |

5、permission

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 字段名称 | 数据类型 | 长度 | 允许为空 | 中文说明 | 备注 |
| permissionId | Varchar | 32 | 否 | 权限编号 | 主键 |
| perName | Varchar | 20 | 否 | 权限名称 |  |
| description | Varchar | 45 | 否 | 权限描述 |  |
| isAvailable | boolean |  | 否 | 权限可用 | 默认true |

1、algorithm

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 字段名称 |  | 数据类型 | 长度 | 允许为空 | 中文说明 | 备注 |
| algorithmId |  | Varchar | 32 | 否 | 算法编号 | 主键 |
| type |  | Varchar | 20 | 否 | 算法类型 |  |
| interfaceDesc |  | Varchar | 45 | 是 | 接口描述 | 默认为空 |

2、algorithm\_para

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 字段名称 | 数据类型 | 长度 | 允许为空 | 中文说明 | 备注 |
| algorithmId | Varchar | 32 | 否 | 算法编号 | 外键 |
| paramId | int | 11 | 否 | 参数编号 | 主键 |
| dataType | Varchar | 20 | 否 | 数据类型 |  |
| paramDesc | Varchar | 100 | 是 | 参数描述 | 默认为空 |
| isNecessary | boolean |  | 是 | 是否可选 | 默认为空 |

实践任务：

1. data\_mining\_task

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 字段名称 | 数据类型 | 长度 | 允许为空 | 中文说明 | 备注 |
| taskId | Varchar | 32 | 否 | 任务编号 | 主键 |
| type | Varchar | 20 | 否 | 任务种类 |  |
| taskDesc | Varchar | 45 | 是 | 任务描述 | 默认为空 |
| finishTime | Date |  | 是 | 截止时间 | 默认为空 |
| startTime | Date |  | 是 | 开始时间 | 默认为空 |
| duration | Varchar | 45 | 是 | 持续时间 | 默认为空 |

1. data\_mining\_result

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 字段名称 | 数据类型 | 长度 | 允许为空 | 中文说明 | 备注 |
| resultId | Varchar | 32 | 否 | 结果编号 | 主键 |
| type | Varchar | 20 | 否 | 结果种类 |  |

3、datasets\_assigned

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 字段名称 | 数据类型 | 长度 | 允许为空 | 中文说明 | 备注 |
| taskId | Varchar | 32 | 否 | 任务编号 | 主键 |
| collectionId | Varchar | 32 | 否 | 数据集编号 | 主键 |

4、algorithm\_config

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 字段名称 | 数据类型 | 长度 | 允许为空 | 中文说明 | 备注 |
| taskId | Varchar | 32 | 否 | 任务编号 | 主键 |
| algorithmId | Varchar | 32 | 否 | 算法编号 | 主键 |

5、task\_result\_re

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 字段名称 | 数据类型 | 长度 | 允许为空 | 中文说明 | 备注 |
| taskId | Varchar | 32 | 否 | 任务编号 | 主键 |
| resultId | Varchar | 32 | 否 | 结果编号 | 主键 |

6、group\_task\_re

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 字段名称 | 数据类型 | 长度 | 允许为空 | 中文说明 | 备注 |
| taskId | Varchar | 32 | 否 | 任务编号 | 主键 |
| groupId | Varchar | 32 | 否 | 分组编号 | 主键 |

学生：

1、student

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 字段名称 | 数据类型 | 长度 | 允许为空 | 中文说明 | 备注 |
| studentId | Varchar | 32 | 否 | 学号 | 主键 |
| studentName | Varchar | 20 | 否 | 学生姓名 |  |
| classId | int | 11 | 否 | 班级号 |  |
| profession | Varchar | 45 | 否 | 专业 |  |

2、group

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 字段名称 | 数据类型 | 长度 | 允许为空 | 中文说明 | 备注 |
| groupId | Varchar | 32 | 否 | 组号 | 主键 |
| groupLeaderId | Varchar | 32 | 否 | 组长学号 |  |
| groupName | Varchar | 45 | 否 | 小组名 |  |
| setupDate | Date |  | 否 | 设立时间 | 默认为空 |

3、group\_compos\_re

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 字段名称 | 数据类型 | 长度 | 允许为空 | 中文说明 | 备注 |
| studentId | int | 11 | 否 | 学号 | 主键 |
| groupId | int | 11 | 否 | 组号 | 主键 |

数据集管理：

1、provider

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 字段名称 | 数据类型 | 长度 | 允许为空 | 中文说明 | 备注 |
| id | Varchar | 32 | 否 | 提供者 | 主键 |
| name | Varchar | 20 | 否 | 提供者姓名 |  |
| description | Varchar | 45 | 否 | 提供描述 |  |
| email | Varchar | 45 | 否 | 邮箱 |  |

2、data\_set\_collec

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 字段名称 | 数据类型 | 长度 | 允许为空 | 中文说明 | 备注 |
| id | int | 11 | 否 | 数据库编号 | 主键 |
| characteristics | Varchar | 20 | 否 | 特征 |  |
| enableMissing | Varchar | 45 | 否 | 允许缺省 |  |
| description | Varchar | 45 | 否 | 描述 |  |
| dateDonated | Varchar | 45 | 否 | 捐赠时间 |  |
| relevantePapers | Varchar | 255 | 是 | 相关论文 |  |
| abstract | Varchar | 255 | 否 | 描述摘要 |  |
| associatedTasks | Varchar | 255 | 是 | 任务相关 |  |
| topic | Varchar | 100 | 是 | 数据主题 |  |
| area | Varchar | 20 | 是 | 来自地区 |  |

# 模块设计

## 详细模块设计

### 登录注册模块

**1、在线注册**

在登录之前，未拥有账号的学生可以选择创建新账号，采用学号注册。注册时需要输入账号（学号）、姓名、密码和再次确认密码，提交后系统会对输入数据的有效性（例如输入项是否符合格式、两次密码是否一致、学号是否已存在）进行验证，若符合条件则注册成功，否则提示重新注册，返回注册界面。

**2、用户登录**

学生和管理员统一由此入口登录，学生输入学号和密码，管理员（即教师）输入工号和密码。输入的账号跟密码会验证是否合法。

输入登录所需要的账号密码后，提交登录信息，后台的权限管理框架会验证提交信息的有效性，返回给前台一个令牌。所有用户的访问需要携带给令牌访问系统，令牌存在有效期。任何对于系统背后的模块访问（即使url暴露），都需要验证用户授权信息，无权限的访问操作将被重定向到登录页面。则进入系统内部，否则提示出错信息，返回登录界面。

### 数据集管理模块

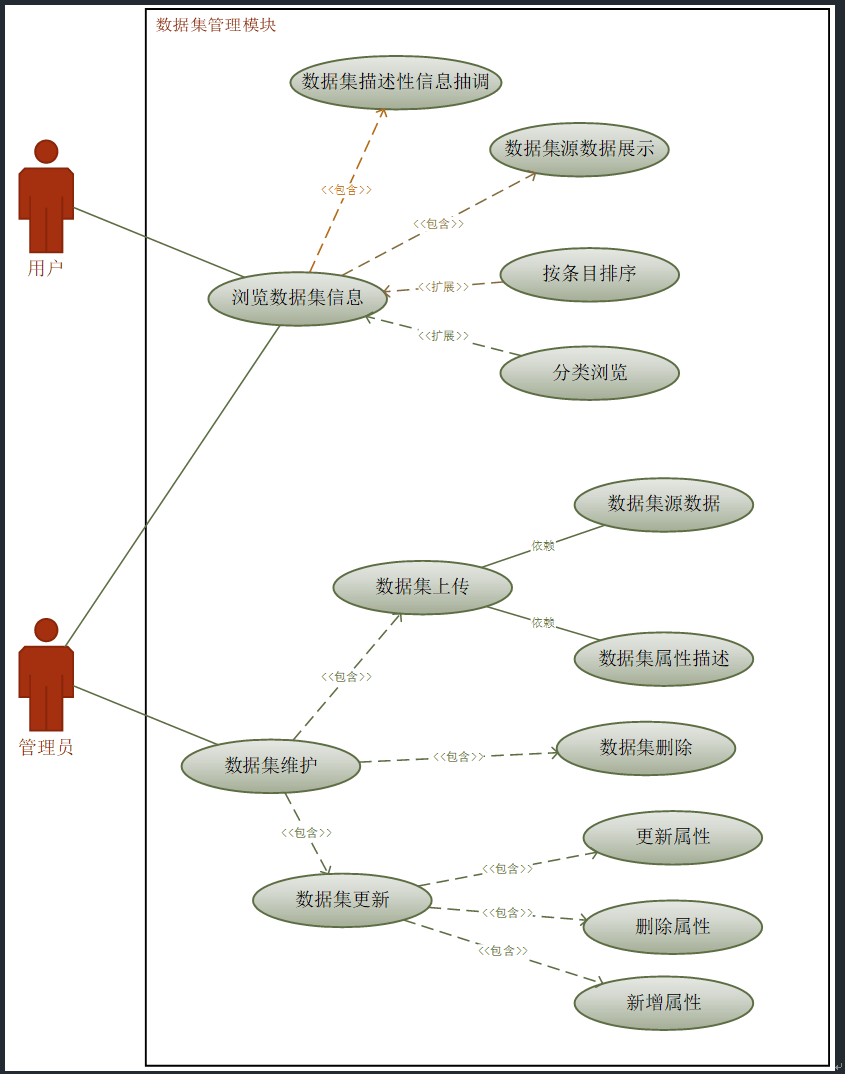
****

图 5.1数据集管理模块用例图

**1、数据集概况展示**

该板块的首页是展示数据集的概况，学生和管理员都可通过从首页了解到该管理系统中所有数据集的数据类型、数据数量、数据挖掘形式等概括性信息，从总体上给用户直观的印象。

该板块支持数据集库根据不同的分类进行排序查看，即用户可以在表中点击不同的基本信息，数据集库就会根据内容的字母或数字等顺序，进行排序展示数据集库。

同时该板块也支持分类浏览数据集信息，用户可以通过点击分类框内的不同分类内容，查看所需要的类型的数据集。用户也可以通过表格查询关键字的方式直接对所需要的类型进行查询筛选。

**2、数据集具体展示**

数据集的具体内容是由首页的数据集库选择进入单个数据集后展示的，它是数据集内容的具体呈现，包含了数据集的基本描述信息以及具体内容，同时可以在此页面进行数据集的。

基本描述信息包括数据集的名称、年份、实例数、所属的类型（分类、聚类）等，数据集的具体内容通过表格的形式呈现，支持用户进行查询和排序的简单操作，并且能够按照不同的条目（如年份、类型）进行分类排序浏览。对于管理员而言，在该页面中还支持对数据集的基本信息及数据集内的具体属性进行增加、删除。

**3、数据集信息管理**

对于学生而言，对数据集可以使用的操作仅有下载，学生可以根据自己的任务所需数据集，到该数据集所在页面，下载数据集及其描述文件，进行数据挖掘的具体操作。

而对于管理员而言，对数据集的处理包括数据集的上传、删除、更新及下载。上传数据集是提供一个单独页面进行数据集的上传，以及对于数据集描述信息进行填写，由后台保存和读入。删除则是通过表单选择需要删除的数据集，直接对数据集进行删除，删除的同时会提示相应的删除信息，如已被关联的数据集删除警告、数据量大的评估警告。更新则是可以对数据集的基本信息，以及数据集内的具体属性进行增加删除，具体是在各个数据集页面中实现。

### 学生信息管理模块

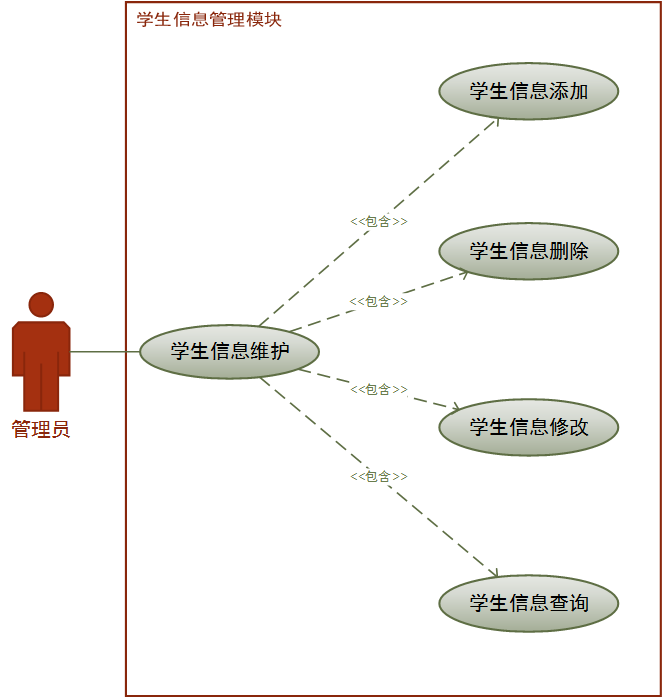


图5.1.3 学生信息管理模块用例图

**1、学生概况展示**

该模块主要是管理员对于学生信息的管理，其首页主要展示学生的信息概况，包括系统内已记录的学生的信息数、一天内新增学生数量、学生增加量趋势等一些数据信息。

**2、学生信息管理**

管理员（教师）可以对学生的信息进行添加、修改、删除和查询，具体的学生信息包括学生的姓名、学号、登录密码、班号、专业等，教师可以根据自己的需求进行修改。

### 实践任务管理模块

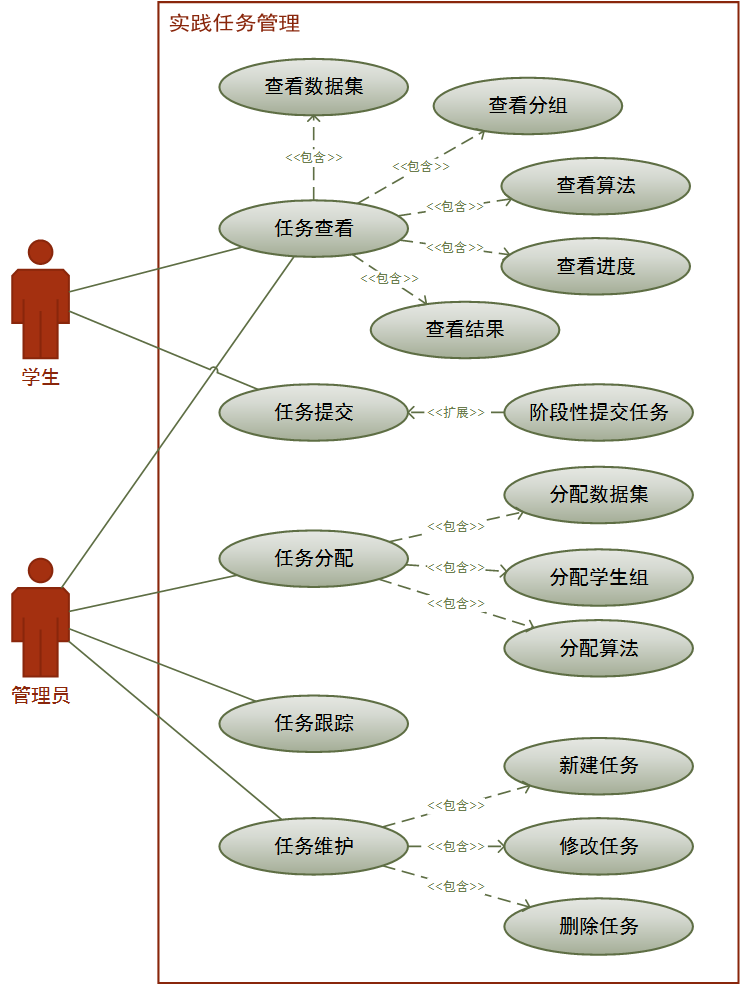


图5.1.4 实践任务管理模块用例图

**1、任务情况展示**

该模块主要是数据挖掘任务的管理和分配，对于学生而言，主要是任务分配后情况的展示，主要呈现的自己所执行的任务、使用的数据集、以及使用的一个或多个算法、执行进度。通过任务情况的展示，学生可以了解自己所执行的任务。

**2、提交执行任务**

学生可以根据自己所执行的任务的进度，进行阶段性提交总结，任务可以多次提交，查看具体结果，也可以对已提交的结果进行查看。

**3、后台任务分配**

系统会根据历史的任务状况自动筛选合适的学生，运行默认的算法配置好任务，分配数据集库内的数据集，并为任务合理指定几组不同的学生。

每组学生每次只能分配一组任务，一个任务可以由多组学生参与，同时也配置相对应的几个算法。

同时除了系统默认生成任务配置，可由管理员执行偏好配置，学生当前的任务状态可在系统表现出来，对于任务冲突等学生被分配任务给予操作提示。

**4、管理修改任务**

管理员可以根据任务分配的情况进行修改。包括新增挖掘任务、对任务所分配的数据集定向修改、自由调整学生的分组状态、任务算法的。

**5、任务跟踪分析**

系统可以记录下学生执行数据挖掘任务的进度，呈现学生任务的阶段。同时可以调用学生完成任务后得到的结果存储记录，比较结果算法、算法参数、实验方法等因素造成的差异。

### 数据挖掘算法库管理部分

这个模块主要是对数据挖掘算法的提供和展示，支持算法（分类、聚类等）的分类浏览，可以对具体算法进行阅读和浏览，浏览包括算法参数、类型、接口的综合描述。

本系统的算法在底层默认实现，暂不提供对外的算法导入，以便统一在本平台进行数据发掘任务的接口与参数。