数据挖掘实践教学平台研究与实现 需求分析

# 引言

* 1. **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，

* 1. **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。

* 1. **数据挖掘概览**

1.3.1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。

1.3.2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：

（1）数据清洗：

数据质量问题：

* 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：

* 数据的完整性：

1. 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；
2. 通过前后数据补全：如通过前后时间均值来补全缺失时间；
3. 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：

1. 按主键去整理：保留唯一数据；
2. 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：

1. 值域合法性：设定数值上下界来确保数据值的合法性；
2. 值类型合法性：如日期格式必须为2017/5/6；

* 数据的一致性:

（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：

（1）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；

（2）测试方案设计：设计某种测试模型的质量和有效性的机制；

（3）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；

（4）模型测试评估：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：

（1）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；

（2）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；

（3）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。

# 功能模块



2.1数据集管理模块

2.1.1功能描述：

（1）数据集展示：展示数据集的名称、种类、实例数、主题、属性等基本信息；

（2）数据集浏览：浏览数据集具体的属性、来源具体描述，亦可展开具体的源数据；

（3）数据集下载：提供用户对外的数据集的下载功能；

（4）数据集维护：给拥有维护权限的管理员上传、删除、更新数据集的功能；

