**《大数据与数据挖掘》课程教学大纲**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 英文课程名 | Big Data and Data Mining | | 总 学 时 | | 32 | 学 分 | 2 |
| 课程编码 | G126156 | | 理论教学学时 | | 24 | 适用专业 | 软件工程 |
| 课程类别（请在课程所属类别栏注明选修或必修） | 通识课程 |  | 实践  教学  学时 | 实验学时 | 0 | 先修课程 | 离散数学，数据结构，概率论，数据库系统原理，计算机程序设计 |
| 大类基础课程 |  | 上机学时 | 8 | 开课学院（部） | 计算机学院 |
| 专业课程 | 专业选修课 | 其它 | 0 | 基层教学组织 | 操作系统课程群教学团队 |

**一、课程简介**

大数据管理的一个重要工具就是数据仓库，数据仓库是一个面向主题的、集成的、相对稳定的、反映历史变化的数据集合，用于支持管理决策；数据挖掘又称为数据库中的知识发现，是基于人工智能、机器学习、统计学等技术，高度自动化地分析原有的数据，进行归纳性推理，从数据仓库或数据库中提取可信的、新颖的、有效的、人们感兴趣的、能被人理解的知识的高级处理过程。这些知识是隐含的、事先未知的有用信息，提取的知识表现为概念、规则、模式、规律等形式，以帮助管理者做出正确的决策。数据挖掘是一门新兴的交叉性学科，是在信息技术领域迅速兴起的计算机技术。数据挖掘技术面向应用。在很多重要的领域，数据挖掘都发挥着积极的作用。MIT技术评论在2002年将数据挖掘评为影响今后技术发展的十大技术之一，而在2013年则将深度学习列为突破性科学技术的榜首。因此，这门课程是物联网专业及相关专业的重要课程之一。

**二、教学目标**

**2.1 课程教学目标**

本课程的目的：

1. 使学生掌握数据挖掘的基本概念、相关技术，了解数据挖掘在数据处理和规则提取中的应用现状、应用前景和研究方向；
2. 强化在数据挖掘中的算法设计、分析的能力，同时通过对经典算法的实现、与已有实现的性能比较，培养学生软件开发能力以及创新能力；
3. 通过对一些主题的调研，培养学生阅读外文资料、文献调研方面的能力。

**2.2 课程目标与毕业要求（指标点）对应关系**

该课程支撑以下毕业要求：

【毕业要求1】工程知识：能够将数学、自然科学、工程基础和专业知识用于解决复杂软件工程问题。

【毕业要求2】问题分析：能够应用数学、自然科学和工程科学的基本原理，识别、表达、并通过文献研究分析复杂软件工程问题，以获得有效结论。

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 课程目标 | 毕业要求指标点 | 教学环节 | | | |
| 课堂授课 | 实验 | 作业 | 课堂讨论 |
| **目标1**：使学生掌握数据挖掘的基本概念、相关技术，了解数据挖掘在数据处理和规则提取中的应用现状、应用前景和研究方向； | 问题分析：能够应用数学、自然科学和工程科学的基本原理，识别、表达、并通过文献研究分析复杂软件工程问题，以获得有效结论。 | √ | √ |  | √ |
| **目标2**：强化在数据挖掘中的算法设计、分析的能力，同时通过对经典算法的实现、与已有实现的性能比较，培养学生软件开发能力以及创新能力； | 工程知识：能够将数学、自然科学、工程基础和专业知识用于解决复杂软件工程问题。 | √ | √ |  | √ |
| **目标3**：通过对一些主题的调研，培养学生阅读外文资料、文献调研方面的能力。 | 问题分析：能够应用数学、自然科学和工程科学的基本原理，识别、表达、并通过文献研究分析复杂软件工程问题，以获得有效结论。 | √ |  |  | √ |

**三、课程教学内容及学时分配**

**1．理论教学安排**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 序号 | 章节或知识点(模块) | 教学内容 | 学时分配 | 教学要求  (应明确教学重点、难点和教学方法) | 学生任务 | |
| 作业要求 | 其他要求(自学/讨论） |
| 1 | 引言 | 数据仓库、数据挖掘的发展  数据仓库、数据挖掘基本知识，概念与处理步骤  数据仓库、数据挖掘功能  数据仓库、数据挖掘系统的应用  数据仓库、数据挖掘的热点问题及其研究发展方向 | 2 | 讲解数据挖掘的基本概念、研究方向和应用领域，结合自己专业知识，思考数据挖掘具体可以解决什么问题；数据挖掘的基本概念和解决问题的步骤 | 课后阅读：“Top 10 algorithms in data mining” |  |
| 2 | 数据预处理 | 描述性数据汇总  数据清理  数据集成  数据变换  数据归约  数据离散化  概念分层生成 | 4 | 掌握数据预处理的一些重要方法，了解最新研究进展和方向。举一个实际的问题，学生来尝试解决。 | 查阅文献选一篇或多篇熟悉并讨论 | 讨论Top10中感兴趣的一两种算法 |
| 3 | 数据仓库导论 | 数据仓库概念  概念分层生成  数据立方体  OLAP | 4 | 掌握数据仓库一些重要概念 | 阅读：Inmon“Building the Data Warehouse” |  |
| 4 | 关联规则 | 购物篮中的秘密  关联规则：基本概念  Apriori算法  AprioriTID算法  FP-Growth算法  多层关联规则 | 4 | 重点掌握关联规则算法的一些重要概念以及一到两种关联规则算法，安排一次上机实践，实现一种关联规则挖掘算法 | 阅读Rakesh的开创性文献 | 讨论数据预处理的文献阅读新的 |
| 5 | 分类与预测 | 决策树归纳  贝叶斯分类  基于规则分类  其它分类法（后向传播分类、SVM、KNN、CBR、关联规则分类）  预测、评估、集成方法介绍、模型选择 | 4 | 重点掌握分类算法的一些重要概念以及一到两种分类算法，安排一次上机实践，实现一种决策树分类算法 | 阅读ID3、C4.5文献 | 讨论关联规则的文献阅读心得和实现心得 |
| 6 | 聚类分析 | 聚类方法概述  基于划分的算法  K-Means聚类  层次聚类算法  其它算法 | 4 | 安重点掌握聚类算法的一些重要概念以及一到两种聚类算法，安排一次上机实践，实现一种聚类算法 | 准备聚类算法调研、讨论 | 讨论分类的文献阅读心得和实现心得 |
| 7 | 数据挖掘前沿 | 文本处理和分析 | 2 | 掌握和了解数据挖掘的最新研究动态和发展方向 |  | 讨论聚类的文献阅读心得和实现心得 |

**2．实践教学安排**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 序号 | 项目名称 | 学时 | 类型 | 每组人数 | 教学要求  (应明确教学重点、难点和教学方法) | 学生任务 | |
| 作业要求 | 其他要求(自学/讨论) |
| 1 | 关联规则算法 | 4 | 验证设计 | 1 | 通过关联规则算法、分类算法、聚类算法实现，掌握数据挖掘的流程：数据准备、预处理、挖掘、结果、评测 | 阅读关联规则算法 | 自学编程语言和编程环境 |
| 2 | 分类算法 | 2 | 验证设计 | 1 | 阅读分类算法 |
| 3 | 聚类算法 | 2 | 验证设计 | 1 | 阅读聚类算法 |

**四、考核方式及成绩评定方式**

本课程为考查课，课程重点在学生掌握数据挖掘算法的实现技术，掌握算法评价机制及其机器学习模型选择、评估、验证方法。平时的学习较为重要，对一些经典问题的算法调研、实现、比较，并对实验结果进行分析，培养文献获取、软件开发的能力，进程性成绩占50％；期末上交一份数据挖掘算法实现与应用的综合报告，占50％。

**五、教材、课程网址及参考书目**

教材：黄德才. 数据仓库与数据挖掘教程. 清华大学出版社; 第1版 (2016年8月)

参考书：

**[1]** 韩家炜，M. Kamber.《数据挖掘：概念与技术》. 机械工业出版社; 第3版 (2014年11月).

**[2]** Jiawei Han, Micheline Kamber, Jian Pei. Data Mining: Concepts and Techniques. 机械工业出版社; 3rd Edition (2012年3月).

**执笔者：范玉雷**

**审核者：**

**课程教学团队成员：杨良怀、黄德才、龚卫华、梅建萍、范玉雷、高楠**