《数据挖掘》课程教学大纲

课程编号：	06302525	**课程性质：**	专业选修
课程名称：	数据挖掘	**学时/ 学分：**	32+16/2.5
英文名称：	Data Mining	**考核方式：**	期末笔试
选用教材：	《数据挖掘—概念与技术》（第三版）Jiawei Han 等著，范明等译， 机械工业出版社	**大纲执笔人：**	张静
先修课程：	高等数学	**大纲审核人：**	陈志华
适用专业：	计算机		

一、教学基本目标

《数据挖掘》是计算机专业本科生的一门的专业选修课，旨在使学生掌握数据挖掘的基本概念、原理和技术，了解数据挖掘的产生背景、研究意义，深入学习数据挖掘的主要内容、功能和技术，对数据仓库、概念描述、关联规则挖掘、分类、聚类等内容进行深入学习，掌握各种数据挖掘技术的代表性算法的原理及实现。要求学生通过本门课的学习，基本掌握数据仓库与 OLAP、数据预处理、关联规则挖掘、分类、聚类等数据挖掘的主要技术，初步了解数据挖掘的其他相关技术，数据挖掘的应用及发展趋势，同时对数据挖掘这门综合了数据库技术、统计学、机器学习等多个学科的交叉学科有一个总体的了解和较为深入的认识。该门课程培养了学生对大规模数据的分析和理解能力，为解决社会实际问题的毕业能力达成提供支持。

二、课程涉及知识技能

本课程要求学生能够应用数学、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工程问题，以获得有效结论，能够设计针对复杂工程问题的解决方案，设计满足特定需求的算法，并能够在设计环节中体现创新意识，考虑算法的有效性及社会效应。能够基于科学原理并采用科学方法对复杂工程问题进行研究，具有针对相应的数据挖掘问题，设计算法，完成实验及分析，并能够解释数据、通过综合实验结果得到合理有效的结论。

三、相关能力培养

本课程为计算机科学技术专业的专业选修课，通过该课程的学习，学生应具备基于工程相关背景知识进行合理分析的能力，具备复杂工程问题和预测和模拟能力，工程问题的社会分析能力，以及理解应承担的社会责任。通过该课程的学习，学生应具有自主学习和终身学习的意识，具有不断学习和适应发展的能力。

四、教学基本内容

各章节详细介绍：

（一） 引言

主要介绍数据挖掘的基本概念、数据挖掘的对象、数据挖掘的功能、数据挖掘的分类等内容，最后给出数据挖掘面临的主要问题，及发展的导向。

重点掌握：数据挖掘的概念，数据挖掘的对象及功能。

（二） 认识数据

本章主要介绍数据挖掘面对的数据对象与属性类型，数据的基本统计描述，数据可视化，以及度量数据的相似性和相异性的方法等。

重点掌握：数据挖掘属性类型，数据的统计描述方法，相似性计算方法。

（三） 数据仓库与联机分析

本章从数据仓库的概念开始，介绍了多维数据模型（星型、雪花、事实星座）、数据立方体、概念分层、OLAP 等概念和原理，接下来对数据仓库的实现做了较为详细的介绍，最后给出数据立方体技术的进一步发展以及从数据仓库到数据挖掘的发展演变过程。

重点掌握：多维数据库模式，数据立方体、OLAP。

（四） 数据预处理

数据预处理是数据挖掘的前期工作，是数据挖掘的重要组成部分。本章首先介绍数据预处理的背景，然后详细地介绍数据预处理的主要内容，包括数据清理、数据集成和变换、数据规约、离散化和概念分层生成等。

重点掌握：数据清理、集成和变换的算法，包括：分箱、最小-最大规范化、z-score 规范化。

难点：数据变换，数值规约。

（五） 概念描述

概念描述是描述性数据挖掘的主要类型。本章主要介绍概念描述的原理、数据概化和基于汇总的特征化（属性相关分析）、挖掘类比较（区分不同的类）、在大型数据库中挖掘描述统计度量等内容。

重点掌握： 属性相关分析，类比较方法。

（六） 挖掘频繁模式、关联和相关性

关联规则挖掘是数据挖掘的重要组成部分，本章首先介绍关联规则的基本概念，关联规则的分类、接下来重点介绍两个最主要的关联规则算法：Apriori 算法和 FP-tree 算法。最后简单介绍一下冰山查询，以及基于约束的关联规则挖掘。

重点掌握： Apriori 算法，FP-tree 算法。

难点：Apriori 算法，FP-tree 算法。

（七） 分类

分类和预测是重要的数据挖掘功能，本章首先介绍分类和预测的概念，以及一些基本概念，接下来重点介绍几种重要的分类算法，包括：判定树分类、贝叶斯分类、后向传播分类等，然后对预测算法进行介绍，包括：线性回归、多元回归、非线性回归等。最后给出分类的评估标准。

重点掌握内容：判定树归纳、朴素贝叶斯分类、贝叶斯信念网络、多层前馈神经网络、k-近邻分类、线性回归、非线性回归。

难点：贝叶斯信念网络、多层前馈神经网络、非线性回归

（八） 聚类分析

聚类分析是数据挖掘的重要功能之一，本章将详细介绍聚类分析的概念，聚类分析中的数据类型，主要聚类方法的分类，重点介绍几类聚类方法，包括：基于划分的方法、基于层次的方法、基于密度的方法，基于网格的方法，基于模型的方法等，最后给出孤立点分析的方法。

重点掌握内容：基于划分、基于层次、基于密度、基于网格的聚类方法，独立点分析。

难点：BIRCH 算法、Chameleon 算法、DBSCAN 算法、DENCLUE 算法、STING 算法。

实验内容介绍：

实验 1　Apriori 算法研究与实现　利用 C++/Java/python/matlab 实现 Apriori 算法，并在给定数据上完成实验及实验结果分析。

实验 2　基于 SVM 的数据分类　利用 C++/Java/python/matlab 实现 SVM 分类算法，并在给定数据上完成实验及实验结果分析。

实验 3　基于划分的聚类算法研究与实现　利用 C++/Java/python/matlab 实现 K 均值和 K 中心点聚类算法，并在给定数据上完成实验及实验结果分析。

五、建议教学进度

本课程讲课需要 32 课时，实验需要 16 课时。

(一) 引言　　授课：3 课时

(二) 认识数据　授课：4 课时　　自学：2 课时

(三) 数据预处理　　授课：4 课时　　自学：2 课时

(四) 数据仓库与联机分析处理　　授课：3 课时　　自学：2 课时

(五) 概念描述　　授课：2 课时　　自学：2 课时

(六) 挖掘频繁模式、关联和相关性　　授课：4 课时　　自学：2 课时

(七) 分类　　授课：6 课时　　自学：6 课时

(八) 聚类分析　　授课：6 课时　　自学：6 课时

(九) 实验一（6 课时）

(十) 实验二（4 课时）

(十一)　实验三（6 课时）

六、教学方法

数据挖掘是一门理论性很强的数据库前沿知识课程，该课程的内容涉及数据库技术、人工智能、机器学习、神经网络、统计学、模式识别、知识库系统等多学科多领域。鉴于该课程学习内容繁多，知识体系庞杂，该课程的学习主要以知识传授为主，以教师为主导，系统讲授数据挖掘的原理、技术、功能、算法等，

对该课程涉及到的庞大知识体系，抓重点、难点、关键，精讲算法的原理和实现，使学生深入了解数据挖掘的完整知识结构。在知识传授阶段，采用算法原理讲解与练习相结合的方式，使学生通过自己编写算法的实现代码，深入理解算法的本质。课堂采用教师讲述与学生小组讨论相结合的方法，对重点难点算法，进行深入讲解，并要求学生通过实践练习，加深对算法的认识，培养学生主动思考和动手编程实践的能力。最后根据出勤、平时作业、实验成绩以及期末笔试等 4 项评定考核成绩。

七、考核方式

期末闭卷笔试。

八、成绩评定方法

考核以百分制记分，总成绩由考勤、作业及实验成绩（30%）、期末笔试成绩（占 70%）构成。

九、教学参考书

- 《Data Mining: Practical Machine Learning Tools and Techniques》Second Edition. Lan H. Witten, Eibe Frank, Morgan Kaufmann. 2011。
- 《Introduction to Data Mining》Pang-Ning Tan, Michael Steinbach, Vipin Kumar, 人民邮电出版社（影印版）， 2010。
- 《数据挖掘原理与算法》，毛国君等，清华大学出版社，2009.