《数据挖掘》教学大纲

1．大纲文本

一．课程内容

数据挖掘是从数据集中识别出有效的、新颖的、潜在有用的，以及最终可理解的模式的非平凡过程。知识发现将信息变为知识，从数据矿山中找到蕴藏的知识金块，将为知识创新和知识经济的发展作出贡献。 本课程全面而又系统地介绍了知识发现的方法和技术，反映了当前知识发现研究的最新成果。

二．课外作业

以教材中每章所附的习题为主

三．实验

实验一　关联规则算法(Apriori 算法)

内容：利用关联规则算法，挖掘关联知识。

目的：了解关联规则、频繁集、置信度、支持度的概念。

实验二　分类算法(C4.5 算法、ID3 算法)

内容：　程序实现 C4.5 或 ID3 算法

目的：了解信息熵的概念，掌握算法的基本设计框架。。

实验三　聚类（K-means）

内容：　程序实现 K-means 算法。

目的：　了解距离、相似度等概念，掌握聚类算法的应用

实验四　神经网络分类（感知器算法）

内容：　使用 MatLab 实现多分类

目的：　掌握神经网络的基本原理，熟悉神经网络的应用

实验五　遗传算法的优化（SGA 算法）

内容：使用 C++.net 设计遗传算法解决复杂函数的最优解问题。

目的：初步掌握遗传算法的概念

实验要求：以上实验，根据情况尽可能多的完成，至少选择 2~3 个实验完成。

四．主要参考书

史忠植著，《知识发现》　清华大学出版社　2002.1

各个学术刊物上的最新论文。

2．大纲说明

一．课程的目的和任务

面对日益庞大的数据资源，人们迫切需要强有力的工具来“挖掘”其中的有用信息，数据挖掘就是针对这一需求而发展起来的一门汇集统计学、机器学习、数据库、人工智能等学科内容的新兴的交叉学科，本课程深入探讨数据挖掘原理，把信息科学、计算科学和统计学对数据挖掘的贡献融合在一起，培养计算机专业高年级

本科学生具备初步的科研能力和创造能力。

二．本课程的要求

通过本课程的学习，要求学生初步掌握数据挖掘的重要概念和任务、数据挖掘中的常用算法（决策树、关联规则、范例推理、模糊聚类法、粗糙集、贝叶斯网络、支持向量机、隐马尔科夫模型、进化和遗传算法、神经网络），以及数据挖掘当前的研究动向。

三．本课程与其它课程的联系

本课程的是计算机专业的一门专业课程。学生在学习本课程之前应当具备《高等数学》、《线性代数》、《概率统计》、《程序设计语言》、《数据库原理》等方面的预备知识。

四．各章主要讲解内容

第 1 章 绪论

1.1 知识

1.2 知识发现

1.3 知识发现的任务

1.4 知识发现的方法

1.5 知识发现的对象

1.5.1 数据库

1.6 知识发现与创新

第 2 章 决策树

2.1 归纳学习

2.2 决策树学习

2.3 CLS 学习算法

2.4 ID3 学习算法

2.5 决策树的改进算法

2.6 决策树的评价

2.7 简化决策树

2.8 连续型属性离散化

2.9 基于偏置变换的决策树学习算法 BSDT

2.10 归纳学习中的问题

第 3 章 关联规则

3.1 关联规则挖掘概述

3.2 广义模糊关联规则的挖掘

3.3 挖掘关联规则的数组方法

3.4 任意多表间关联规则的并行挖掘

3.5 基于分布式系统的关联规则挖掘算法

3.6 词性标注规则的挖掘算法与应用

第 4 章 基于范例的推理

4.1 概述

4.2 过程模型
4.3 范例的表示
4.4 范例的索引
4.5 范例的检索
4.6 相似性关系
4.7 范例的复用
4.8 范例的保存
4.9 基于例示的学习
4.10 范例工程
4.11 范例约简算法
第 5 章 模糊聚类
5.1 概述
5.2 传递闭包法
5.3 FCMBP 聚类法
5.4 系统聚类法
5.5 C－均值聚类法
5.6 聚类有效性
5.7 聚类方法的比较
第 6 章 粗糙集
6.1 概述
6.2 知识的约简
6.3 决策逻辑
6.4 决策表的约简
6.5 粗糙集的扩展模型
6.6 粗糙集的实验系统
6.7 粗糙集的展望
第 7 章 贝叶斯网络
7.1 概述
7.2 贝叶斯概率基础
7.3 贝叶斯学习理论
7.4 简单贝叶斯学习模型
7.5 贝叶斯网络的建造
7.6 贝叶斯潜在语义模型
7.7 半监督文本挖掘算法
第 8 章 支持向量机
8.1 统计学习问题
8.2 学习过程的一致性
8.3 结构风险最小归纳原理
8.4 支持向量机
8.5 核函数

8.6 基于分类超曲面的海量数据分类方法
第 9 章 隐马尔科夫模型
9.1 马尔科夫过程
9.2 隐马尔科夫模型
9.3 似然概率和前反向算法
9.4 学习算法
9.5 基于状态驻留时间的分段概率模型
第 10 章 神经网络
10.1 概述
10.2 人工神经元及感知机模型
10.3 前向神经网络
10.4 径向基函数神经网络
10.5 反馈神经网络
10.6 随机神经网络
10.7 自组织特征映射神经网络
第 11 章 进化和遗传算法
11.1 概述
11.2 基本遗传算法
11.3 遗传算法的数学理论
11.4 遗传算法的基本实现技术
11.5 遗传算法的高级实现技术
11.6 并行遗传算法
11.7 遗传算法应用
第 12 章 知识发现平台 MSMiner
12.1 概述
12.2 数据仓库
12.3 MSMiner 的体系结构
12.4 元数据管理
12.5 数据仓库管理器
12.6 算法库管理
12.7 数据挖掘任务规划
12.8 关系数据库知识发现查询语言 KDSQI
第 13 章 Web 知识发现
13.1 概述
13.2 Web 知识发现的任务
13.3 Web 知识发现方法
13.4 模型质量评价
13.5 文本分析功能
13.6 文本特征的提取
13.7 基于文本挖掘的汉语词性自动标注研究

13.8 文本分类
13.9 文本聚类
13.10 文本摘要
13.11 用户兴趣挖掘
第 14 章 生物信息知识发现
14.1 概述
14.2 基因的基本结构
14.3 生物信息数据库与查询
14.4 序列比对
14.5 核酸与蛋白质结构和功能的预测分析
14.6 基因组序列信息分析
14.7 功能基因组相关信息分析
14.8 Internet 资源和公共数据库

五．实验要求

认真完成每个实验，并写出实验报告

六．学时分配表

本课程大纲适用于计算机科学与技术专业，总学时为 48 学时

第 1 章 绪论	2 学时
第 2 章 决策树	4 学时
第 3 章 关联规则	4 学时
第 4 章 基于范例的推理	4 学时
第 5 章 模糊聚类	4 学时
第 6 章 粗糙集	2 学时
第 7 章 贝叶斯网络	4 学时
第 8 章 支持向量机	2 学时
第 9 章 隐马尔科夫模型	2 学时
第 10 章 神经网络	4 学时
第 11 章 进化和遗传算法	4 学时
第 12 章 知识发现平台 MSMiner	2 学时
第 13 章 Web 知识发现	2 学时
第 14 章 生物信息知识发现	2 学时
实验课	6 学时