Vocational Education 职业教育, 2019, 8(1), 6-10
Published Online January 2019 in Hans. http://www.hanspub.org/journal/ve
https://doi.org/10.12677/ve.2019.81002

Hans汉斯

Research on Developing Data Mining Course in Independent Colleges

Dan Li

Tongji Zhejiang Ccollege, Jiaxing Zhejiang
Email: lidan6745@163.com

Received: Nov. 21st, 2018; accepted: Dec. 5th, 2018; published: Dec. 12th, 2018

Abstract

Starting from the general trend of the rise of data science, this paper expounds the importance of offering data mining courses in independent colleges, introduces in detail the main contents and curriculum arrangement of data mining courses in independent colleges, and summarizes the experience and lessons in the course of implementing data mining courses in combination with the characteristics of independent colleges.

Keywords

Data Mining Courses, Independent Colleges

独立院校开设数据挖掘课程的研究

李　丹

同济大学浙江学院，浙江　嘉兴
Email: lidan6745@163.com

收稿日期：2018年11月21日；录用日期：2018年12月5日；发布日期：2018年12月12日

摘　要

本文从当今数据科学兴起的大趋势出发，阐述了独立院校开设数据挖掘课程的重要性，详细介绍了独立院校开展数据挖掘课程的主要内容与课程安排，结合独立院校自身特点，总结了数据挖掘课程实施过程中的经验与教训。

关键词

数据挖掘课程，独立院校

文章引用：李丹. 独立院校开设数据挖掘课程的研究[J]. 职业教育, 2019, 8(1): 6-10. DOI: 10.12677/ve.2019.81002

1. 引言

伴随着互联网时代的深层次变革与大数据时代的到来，数据科学凸显出越来越重要的学术价值与市场价值。特别是在国家经济结构产业升级的大背景下，从海量数据中挖掘出具有战略意义的高价值数据显得尤其重要。数据挖掘技术的主旨便是在表象上相互独立、大熵值的数据中挖掘出潜在的数理规则。数据挖掘技术涉及领域极广，包括基础研究、生物医药、交通、物流、语言学习、金融、管理、3D 打印技术、互联网、通信、工业智能化等各个领域。数据挖掘所涉及的分类、预测、回归、关联与决策算法均为大数据与人工智能领域主流基础算法。无论国内、国外，数据挖掘课程并非针对技术与科研人员，而是针对于业务人员，所以本门课程非常适合独立院校应用型大学学生学习。

2. 独立院校开展数据挖掘课程的重要性

数据挖掘是当今数据科学方向中最活跃、最前沿的区域，是学习大数据与人工智能的基础课程。同时也引起了大量学者进行学习研究，如文献[1] [2] [3] [4] [5]作者各自阐述了数据挖掘算法流程中的细节问题。随国家产业升级战略的推进，数据挖掘工程师、大数据分析师、人工智能工程师市场需求极大，人才匮乏、断档现象较为普遍。此外，现代化企业为了自身生存和发展的需要，迫切希望利用数据科学进行辅助分析与决策。比如，客户深层次的肖像刻画、客户需求分析、设备共享分析、客服共享分析、运力共享分析、智能财务管理等领域都需要借助数据挖掘技术得以实现。对于市场广泛需求，应用型独立院校应该积极建设好数据挖掘课程。开设课程目的是让广大同学了解和掌握数据挖掘的基本概念、基本思想和基本算法。对于实际数据问题，会利用所学知识进行分析处理与决策判断。

3. 独立院数据挖掘课程的主要内容

基于独立院校生源的客观现实，数据挖掘课程的主要内容应包括数据挖掘技术概述与 SPSS Modeler 软件介绍、数据清洗、决策树算法、支持向量机算法、人工神经网络算法、贝叶斯网络算法、聚类算法等。

3.1. 数据挖掘技术概述与 SPSS Modeler 软件介绍

此部分首先要明确数据挖掘技术课程的学科定位为实践学科，授课对象为数据科学相关专业学生，授课的目的是让学生学习数据挖掘相关算法，并通过相关程序与软件对数据进行分析，进而辅助决策。授课内容应主要介绍数据挖掘的产生背景、发展现状、未来趋势以及数据挖掘的基本概念。重点讲解介绍数据挖掘主流软件工具 IBM SPSS Modeler 和 Weka 3.8，特别对 IBM SPSS Modeler 基本操作界面、基本功能选项进行介绍。

3.2. 数据清理

在数据科学实践的过程中，由现场采集的数据往往掺杂大量的噪音，也就是数据中会出现数据的缺失、误差和大熵值情况。特别是会遇到训练数据很少，甚至是无数据可训练的情况。针对于误差数据情况，一般采用的处理方法是用该特征全体数据的均值来填充缺失值。另一种方法是分析出整体数据的概率分布，在该分布意义下，利用分布随机产生的数值填充缺失值。对于误差数据情况，主要采取三种策略来进行处理。第一种策略是客观上承认误差数据的合理性，因为在构建数据模型时，要尊重真实数据，

在合理的误差范围内，承认数据的有效性。第二种策略是严苛筛选法，只要是误差数据，便从数据源中剔除该数据。第三种策略类似于缺失数据的处理方法，应用该数据源所属概率分布的随机数值替代误差值。对于大熵值数据一般依据数据源标签项分类，分类的目的是降低熵值。针对训练数据过少，难以训练生成内部规则的问题，通常有两种处理方法。第一种方法是通过仿真生成数据。通过问题模型数据的现实条件，建立仿真模拟模型，通过仿真系统产生大量数据。第二种方法是通过迁移学习的方法产生数据。如果问题模型与已知模型抽象规则相似，则可利用已知模型进行迁移学习产生数据。

3.3. 决策树算法

决策树算法是在 1980 年代中期伴随着机器学习算法的出现而兴起的可视化、可说明的、基于信息熵理论的数学分类模型。“信息论之父”香农借助概率理论创立了信息熵的概念，熵可以反映数据的杂乱程度，数据类别越多样、越混乱，熵值越大；数据类别越单一、越有序，熵值越小。决策树模型的基本思想是建立树形模型，随着节点的层次越低端，熵值就越小。决策树算法主要包含 ID3.0 算法、C4.5 和 CART 算法。它们的显著区别在于树形模型分支分叉的依据不同。ID3.0 算法的分支依据是信息增益，也就是说，要选取数据特征中信息增益最大者为分叉节点。C4.5 算法为 ID3.0 算法的改进算法，它的分支依据是信息增益率。CART 算法又叫分类回归树，既可以用于分类，又可以用于回归，它的分支依据是 Gini 系数。不同于经济学中的 Gini 系数，这里的 Gini 系数本质上是熵函数的线性逼近。

3.4. 支持向量机算法

支持向量机算法最早是由万普尼克在莫斯科大学的数理统计博士论文中提及的一个雏形思想。万普尼克在苏联解体前一年加入了美国贝尔实验室。早期的支持向量机算法是作为人工神经网络的子算法进行发布和推广的。伴随着九十年代中期，支持向量机在手写体字符识别领域的卓越表现，从而进入了十年的黄金发展期。支持向量机最大的亮点体现在处理非线性离散数据，它将低维空间非线性离散数据映射为高维或无穷维空间线性离散数据。但是它的处理重点并没有放在如何求得转换映射上，而是着重求解高维或无穷维空间的内积值。

3.5. 人工神经网络

人工神经网络算法思想来源于最为朴素的感知机模型。人工神经网络算法的算法步骤一般先将数据各属性进行线性回归，并与压缩函数进行复合转化为逻辑回归，再将损失函数正则化，利用凸分析优化算法求其权重最优值，进而调整线性回归权重。事实上，每个神经元都可以视为一个超平面分类。通过增加神经元的个数与层数就可以达到非线性分类的效果。

3.6. 贝叶斯网络算法

贝叶斯网络算法为概率图学派代表算法之一，拓扑结构为有向无环图，理论依据为概率学中的条件概率、全概率与贝叶斯概率公式。全概率公式表述的是一个事件的发生是由多个事件触发形成的。当要求此事件发生的概率，实质上是求该事件与多个触发事件积的和概率。贝叶斯公式又称逆概率公式，是已知该事件已经发生，而求是哪个触发事件触发的概率。以贝叶斯网络算法模型中最为简单的朴素贝叶斯模型为例，朴素贝叶斯网络模型有一个前提假设，要求数据各项属性是相互独立的。在上述前提下，求此有向无环图的联合概率分布，并由联合概率分布结果求得贝叶斯概率结果。

3.7. 聚类算法

聚类算法为无监督学习算法的代表算法，也就是说在训练模型内在的规则时，无需告知标签项，即

不用告知模型数据的具体类别。聚类算法的主要思想是在某种度量意义下通过判别数据点之间的远近程度来进行分类。聚类算法主要包括 K-Means 算法、K 近邻算法和两步聚类算法等。

4. 独立院校数据挖掘课程安排

针对于独立学院学生的实际情况，数据挖掘课程安排 15 个教学周为宜，每周理论课程 2 课时、上机课程 2 学时。具体课程安排如表 1 所示。

Table 1. Data mining course arrangement
表 1. 数据挖掘课程安排

序号	教学内容	理论课时	上机课时
1	数据挖掘和 SPSS Modeler 概述	2	2
2	数据挖掘的知识形式与算法分类	2	2
3	IBM SPSS Modeler 数据的读入与数据集成	2	2
4	IBM SPSS Modeler 变量的管理与样本管理	2	2
5	挖掘预处理：数据清理	2	2
6	决策树 ID3.0 算法	2	2
7	决策树 C4.5 与 CART 算法	2	2
8	BP 人工神经网络	2	2
9	径向基函数网络	2	2
10	支持向量机	2	2
11	贝叶斯网络	2	2
12	马尔科夫毯网络	2	2
13	K-Means 聚类	2	2
14	Knn 聚类	2	2
15	两步聚类	2	2
	总计	60	

5. 总结

各个独立院校在开展数据挖掘课程的过程中，事实上有很多宝贵的经验可以借鉴。一是授课对象的选择上，较适合选择大学三年级理工科或经济类学生进行教学。这是因为大三学生已经具备一定的编程基础，并且已经学习过高等数学、线性代数和概率论与数理统计，数学理论储备充足，学习数据挖掘技术课程的条件已经具备。二是在算法理论的讲解过程中应循序渐进，逐渐增加难度，不要图快猛进。例如在讲解人工神经网络算法时，应先从最初级的线性回归、逻辑回归、感知机模型开始讲起，再讲如何利用最速下降法或随机最速下降法求解损失函数的最优参数。当然，在数据挖掘课程教学过程中也有很多教训。最为常见的问题便是不注重上机环节，没有抓住数据挖掘课程教实践学科的本质。独立院校只要科学的安排课程内容，明确正确的课程目标，务实的进行上机实践，完全有能力为国家培养大量的数据科学人才。

基金项目

同济大学浙江学院第七届教改项目(项目编号：0118037)。

参考文献

[1] 刘明吉, 王秀峰, 黄亚楼. 数据挖掘中的数据预处理[J]. 计算机科学, 2000, 27(4): 54-57.

[2] 钟晓, 马少平, 张钹, 俞瑞钊. 数据挖掘综述[J]. 模式识别与人工智能, 2001, 14(1): 48-55.

[3] 贺玲, 吴玲达, 蔡益朝. 数据挖掘中的聚类算法综述[J]. 计算机应用研究, 2007, 24(1): 10-13.

[4] 慕春棣, 叶俊. 用于数据挖掘的贝叶斯网络[J]. 软件学报, 2000, 11(5): 660-666.

[5] 李德仁, 王树良, 李德毅, 王新洲. 论空间数据挖掘和知识发现的理论与方法[J]. 武汉大学学报, 2002, 27(3): 221-233.