Advances in Education 教育进展, 2023, 13(3), 1077-1083
Published Online March 2023 in Hans. https://www.hanspub.org/journal/ae
https://doi.org/10.12677/ae.2023.133170

Hans汉斯

基于需求驱动的数据挖掘案例平台的构建

金秀玲

闽江学院数学与数据科学学院，福建 福州

收稿日期：2023年2月12日；录用日期：2023年3月9日；发布日期：2023年3月16日

摘　要

数据挖掘作为本科高校统计学专业的一门专业课程，各校都有自己的考量。但是无非取自不同的教材，用不同的案例对其进行集中的展现。本文就《数据挖掘》课程案例的平台的构建，提出基于需求驱动的案例背景，从而提高数据挖掘课堂教学的时效性。

关键词

数据挖掘，案例教学，需求驱动

Construction of Requirement-Driven Data Mining Case Platform

Xiuling Jin

College of Mathematics and Data Science, Minjiang University, Fuzhou Fujian

Received: Feb. 12th, 2023; accepted: Mar. 9th, 2023; published: Mar. 16th, 2023

Abstract

Data Mining is a professional course of statistics major in colleges and universities, and each school has its own considerations. However, it is nothing more than taken from different teaching materials and concentrated on it with different cases. This paper puts forward a demand-driven case background for the selection of data mining course cases and the construction of the platform, to improve the timeliness of data mining class teaching.

Keywords

Data Mining, Case Teaching, Demand-Driven

文章引用: 金秀玲. 基于需求驱动的数据挖掘案例平台的构建[J]. 教育进展, 2023, 13(3): 1077-1083.
DOI: 10.12677/ae.2023.133170

1. 引言

《数据挖掘》课程作为培养数据挖掘能力的载体，为了达到让学生掌握数据挖掘理论知识以及如何使用统计软件工具来解决实际问题，全面提高学生如何使用数据挖掘这个工具解决实际问题的能力，从而培养数据挖掘应用型的人才。

目前，高校中数据挖掘能力的培养模式中，偏重于数学模型算法的理论知识学习。其中实践案例的选取和实验操作部分存在数据理想化和案例载体更新慢、时效性严重滞后的问题。实验教学内容局限于数据挖掘的理论知识。

对此许多专家学者对这个问题进行探讨，如曾垂省[1]对生物信息学专业的《数据挖掘》课程改革，从教材建设到考核方式提出几点见解；汪一百[2]提出利用网络数据资源对《数据挖掘》课程改革。蔡莉[3]转型经济下资源驱动与机会驱动企业创业的数据挖掘改革；赵晓明[4]从降低理论难度，创建公安专业案例进行《数据挖掘》课程在公安专业的改革；蔡发良[5]以数据思维为核心的教学目标，理论创新与项目为驱动，推进《数据挖掘》课程的改革。李珊珊[6]把教学内容与就业需求相结合，引进横向和纵向案例提出了一般本科院校的《数据挖掘》课程改革。

针对数据挖掘的理论层次与现实社会需求存在较大差距。致使学生学习兴趣低，不利于数据挖掘能力的培养。这些专家分专业进行探讨课程改革探讨，有的提出与相应的科研项目相结合。这些专家都是从实际问题出发，进行数据挖掘教学改革的探讨。

因此，构建需求驱动的数据挖掘教学实践平台模式，真实体现数据挖掘技术在各个交叉学科的实际应用场景，借助现代化教学手段进行多元化的教学。提高学习的内在能动性，达到对数据挖掘技术应用能力的高效的培养模式。

2. 数据挖掘简介

2.1. 数据挖掘

数据挖掘从信息中“淘金”，在大量信息(包括图片数据和文本数据)中挖掘出隐藏的、未知的、对决策具有潜在意义的信息、模型和趋势，利用这些信息的规律构建适合决策信息的模型，提出预测型决策信息的技术、方法和模型。

2.2. 数据挖掘模型知识技术和分析步骤

数据挖掘实际应用中常见的是数据类型是数值数据、图片数据、声音数据、文字数据等。对于非数值数据应先进行转化为数值型数据处理，进而使用数据挖掘进行数据分析。数据挖掘的技术主要包括了回归建模，分类模拟，聚类分析，关联规则、时序性研究、差异检验、智能选择等技术等。

实施数据挖掘的大致包括如下六个过程，详见图 1；第一：明确数据分析的目标，对数据分析的目标的明确了解并设定目标；第二：收集信息，注意建模后的抽样信息、收集过程中注重信息的质量的把控、信息收集过程中注重信息的准确性；第三：数据收集，包括对数据的挖掘、清洗数据、把数据有效的转化到与模型相应的模式和功能；第四：建立模式，对模式的识别、模式的建立、验证模式；第五：模型评估，设定评价指标、多模式比较、模型优选；第六：模型发布，对模型部署与重构。

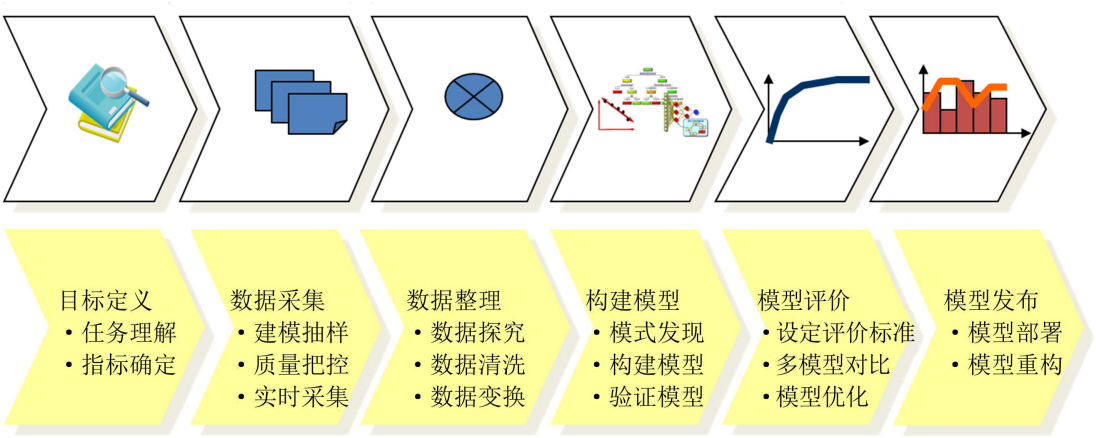

Figure 1. Six steps of data mining
图 1. 数据挖掘六步骤

2.3. 数据挖掘分析能力的等级

数据挖掘中的分析功能，大致表现在如下八个层面，详见图 2：日常报告、即席查询、多维分析、警报、统计分析、预测、可预测性模型设计与优化。

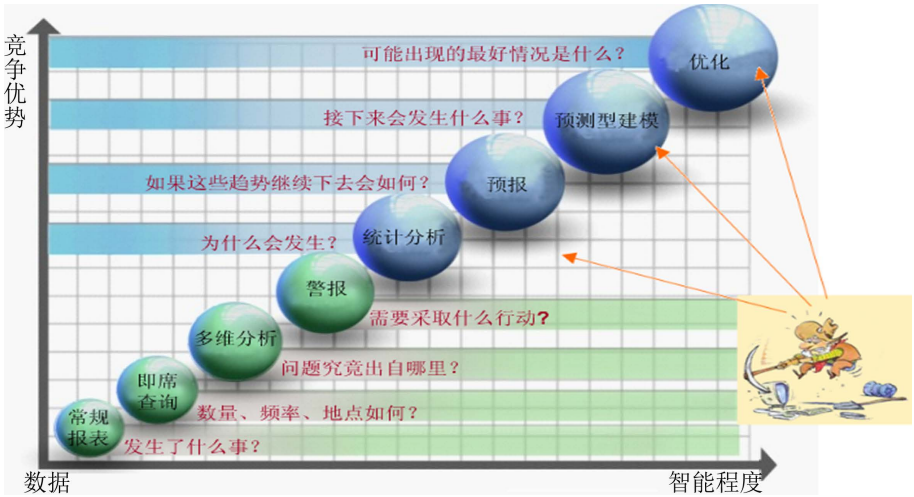

Figure 2. Data analysis ability levels
图 2. 数据分析能力等级

2.4. 挖掘案例实际情况

目前，数据挖掘实践教学中用到的案例普遍存在以下几种情况：第一实验的案例已经载体陈旧，失去的时效性；第二实验的所使用到数据非原始数据、已经是对数据进行过预处理了，该数据直接忽略了数据挖掘的前三个步骤，学生失去了对数据分析锻炼的机会，而且对于在模型的构建一般而言太理想化了，脱离实际应用的场景，学生在以后的无论是参加竞赛还是进行实际的数据分析过程中，会处于拿到数据不知所措、一筹莫展的境地。第三数据分析与相关的专业相脱节。限制了学生对知识面的拓宽，进行学科交叉延展，从而找到自己感兴趣的行业，满足将来就业的需求。

因此，基于现有《数据挖掘》课程课时的局限性和实践案例中的诟病，有必要建立满足企业的需求、社会需要的数据挖掘实践平台。

3. 企业需求的驱动作用

在有限的学时内完成数据挖掘分析能力培养，让学生既要掌握数据挖掘的理论知识还要提高数据分析实际应用能力，使之适合社会之需，能够面向社会和面向市场。深化校企合作合理有效地构建实践平台是一条明确的道路。

3.1. 校企合作必要性

我国的校企合作起步于20世纪70年代，发展于80年代，深入于90年代。校企合作大发展的局面是改革开放以来形成和深化的。校企合作为高等院校的高质量办学和高效益的办学的一项措施。

校企合作的形式也逐步走出校企合作开展专业教育的单一模式，合作面及深度不断拓展。我国校企合作发展不足，与我国工业和高等教育的发展进程、内在机制有着重要的关系。

3.2. 校企合作的优势

首先，通过校企合作满足了社会和市场的需求。校企合作，学校可以根据客户反馈的需求，有针对性地培养技术人才，可以根据市场导向，强调学员的实际能力，也可以培训出市场需求的人才。

再次，校企合作是一个“双赢”的机制。校企合作，就能够实现了高校和企业之间知识、技术人才的资源共享，高校也能够使用了企业所提供的技术设施，而企业也就可以不用再为培训技能人员而担忧大学教育的质量问题，进而达到使学员的在校学习经验和企业实际有机的融合，使高校与企业双方的设备、技能人员等资源达到优势互补，降低了教育质量和企业生产成本，是一个“双赢”的合作方式。

3.3. 学生实践方面

加强模型的技术理论知识的强化，深化专业延展，激发创新力。通过企业真实环境提供的原始数据和实践平台，通过QQ和微信等社交平台实现多渠道的互融互通。一个项目探究解答过程，数据清洗，属性构建，模型核心技术，模型验证和优化实现了完整的数据挖掘的步骤全过程。

3.4. 教师能力培养方面

丰富《数据挖掘》课程教学方式，通过小组内讨论，组间互相提问，相互验证的方式增加学习兴趣。促进教师专业能力提升和双师双能的能力培养。随着数据信息爆炸式的发展，数据挖掘渗透到各行各业，企业的需求弥补了教师掌握社会最前沿的需求以及最新的技术应用。

4. 需求驱动的数据挖掘实践平台

4.1. 对接企事业单位，深化校企合作

引入企业真实的动态业务资源，从真实的社会各个行业需求的实用性角度出发，分析将个性化的需求引入真实试验到《数据挖掘》课程平台，同时结合各行业公开的数据库资源，并能将各种异构资源进行梳理整合在同一平台上，推动了数据挖掘实践案例的开放共享。

1) 与电力部门对接电量使用实时数据，对总体用电量进行预测，估计峰值时间，提出针对性的措施，鼓励居民错峰用电；对用户的窃漏电行为进行挖掘，减少公司的损失。

2) 与电信公司对接，对客户群数据进行梳理，建立优质客户群体。

3) 与大型的水产品养殖企业对接，利用遥感技术收集到的照片，对水质进行检测，判别水质优劣，从而减少人工成本和减低作业风险。

4) 与财政部门对接，收集各类别财政收入数据源，分析影响财政收入的主要因素，预测来年的经济

指标，从而更加科学地作出决策。

5) 与大型商超对接，实现用户的购物习惯构建推荐系统，采用购物篮规则，找到消费者的消费喜好和习惯，建立智能推荐系统，有针对性的发放广告，减少广告支出，增加盈利；以及客户对产品评论和购物体验感进行舆情分析。

4.2. 实践案例平台的构建具体的实验案例安排如表 1

案例一：电力窃漏电用户自主识别系统和用电量峰值预测

(1) 在市场营销信息系统和计量监控信息系统中提取可用的数据变量，并通过分布研究和周期性数据分析的手段对电力信息的统计探索研究。

(2) 从目标需求以及建模的相关需要方面考虑，筛选出需要的特征；然后对该特征采用拉格朗日插值法对缺失值进行插补缺失值和对数据变换重构。构建使用与模型算法的样本指标。

(3) 构建窃漏电用户识别模型；对 CART 决策树模型也进行评价；进行窃漏电诊断。

(4) 构建电量的预测模型；对电量峰值和低谷进行预测，采取相应的措施，让大家错峰使用电器。

在项目研究的进程中，针对统计分析技术的六个环节展开具体的培训与实践。精准提取有用信息、对数据的挖掘，对信息的分布研究与周期性数据分析；包括了样本清洗、缺失值、数据变换，以及建立样本数据的应用。核心计算中神经网络与 CART 决策树构建分类模式的应用；以及模型的评价系统的构建。完整体现了数据分析的 6 个步骤。该案例起到抛砖引玉的作用，将其进行拓展到水用户行为进行分析；以及使用数据挖掘的其他分类算法和预测模型，进行用户行为分析，进行多方位的比较和考量。

案例二：电信公司客户价值分析

(1) 借助电信公司客户数据，对客户进行分类；

(2) 对不同的顾客类别进行特征分析，以比较不同类型顾客的服务能力；

(3) 对不同价值的客户类别提供个性化服务，并制定相应的营销策略。

对精准抽取有效数据的方法、数据探索与分析；数据清洗、缺失值、数据变换等操作；用 RMF 模型构建方法以及 K-Means 聚类方法进行非监督的学习。拓展数据挖掘中处理 RMF 模型和 K-Means 的其他方法分析对用户的情感分析。

案例三：基于水色图像的水质评价

(1) 水质图片数据文件，需要提取图片所属的水质类别，并截取图像中央方块。然后对截取后的图像进行色彩矩阵处理，分为一阶矩阵、二次矩、三级矩阵，并且由于图像存在 R、G 和 B 三种色彩通道，故处理得出的色彩矩阵特征存在九个分量。

(2) 数据分成训练集和测试集；

(3) 结合水质类别和颜色矩特征构成专家样本数据，以水质类比作为目标输出，用训练集构建 LM 神经网络模型。用混淆矩阵和 ROC 曲线测试集评价模型优劣。

图片数据集在打你那光放应用于车牌识别系统，手写输入系统，人脸识别系统，进一步结合遥感技术对极端环境下降雨量的预测识别、根据植物叶子图片构建植物的分类系统，从而开发相应的 APP 达到普及知识的作用。还有其他相关的图片数据上的应用。通过对该案例的学习，以及一系列的流程学习，学生们可以掌握基本图像预处理方法，利用图像处理技术，通过水色图像实现对水质的自动评价并建立 LM 神经网络模型，以及利用混淆矩阵和 ROC 曲线评价模型的优劣。

案例四：财政收入影响因素分析及预测模型

(1) 分析地区财政收入和各级别税收收入情况，研究确定影响地区财政收支的重要特征。通过 Lasso 变量的分析筛选出各级财政收入和各级别税收收入的重要相关因子；

(2) 使用收集的资料形成的已进行了预处理的模型资料，构建 Lasso 的选择模型；

(3) 代入用历史数据训练的神经网络模型，从而对得出了某市的财政收支情况和各种类别的企业未来若干年的财政收入预测值中的预测值。

使用大数据采样，进行信息预处理，同时进行 Lasso 的研究和建立神经网络模式，对某城市地方的财政收入和各级别税收收入做出预报。

案例五：商超产品评论数据情感分析和购物智能推荐系统

(1) 商超评论数据中有众多品牌的评论，提取出某个品牌的评论数据(美的)，并对数据做预处理。

(2) 通过预处理原始数据，并利用 ROSTCM6 对划分评论数据，对文本做分词处理和 LDA 主题分析研究。

(3) 从数据的横向来看，对消费者的购物习惯采用 Apriori 算法进行关系链分析。

对文章信息进行内容处理、对评论文本分类、对文章进行分词处理、对 LDA 问题分析原则的认识和使用，用 LDA 问题分析处理实际问题。拓展到处理 NLP 的卷级神经网络方法进行分析用户情感；文本挖掘推广到社会舆情分析、热门事件舆情分析和影评分析。关联规则可以推广应用到网络黑客的入侵管理、教育系统的管理等。

Table 1. Selection of data mining cases
表 1. 数据挖掘案例的选取

案例	数据预处理方法	数据挖掘核心算法	拓展内容
窃漏电行为的识别	清洗、缺失值、属性变化、构建样本指标	神经网络，决策树	水量预测、水用户行为预测
电信公司用户数梳理	清洗、缺失值、数据变换	RMF 模型，K-Means 聚类	航空系统客户群的梳理
水质图片分类	识别图片数据，转换成 RGB 三阶矩数据，转换数据、构建变量的属性、	LM 神经网络模型	其他分类模型，其他图片识别如人脸识别系统，职务叶子系统等
财政收入影响因素分析	清洗、缺失值、属性变化，变量构建	Lasso 变量选择模型	因素分析模型
商超用户情感分析和购物智能推荐系统	数据清洗、数据转换、文本分词、文本分析构建属性	Apriori 关联规则	爬虫知识、影评、社会舆情分析、网络入侵技术分析、教育数据管理

总之，通过校企合作，在双方共建共赢的战略性目标前提下，建立以需求驱动的《数据挖掘》课程建设的平台构建，达到资源的优化整合，学生真正能做到实验数据的真实性以及接触到社会的各个行业，可以更加明确自己感兴趣的行业和未来自己做数据分析，数据挖掘所要承担的内容。在学校方面，在培养人才方面，少走弯路，直接培养出对社会有用的人才。对企业而言，把新员工的培养放在学校，减少这部分的时间损耗，和培养员工的成本，达到学校、学生、企业三方共赢的有力合作局面。

在实践平台中引入交叉学科真实的数据，构建动态案例资源。把数据挖掘技术切实地得到应用于各个生产部门，如电力、健康医疗、教育行业、经济领域、生产制造业以及公共服务等行业等。利用数据整合技术、元数据技术等将各种异构数据整合在一个统一的网络实践平台上，并结合具体专业和学科的应用开展培训，保障大学生有效地参与数据挖掘能力提升的实际工作。

5. 基于需求驱动的数据挖掘能力培育平台的亮点

(一) 数据挖掘能力培育实践平台的数据来源于交叉学科的真实的数据、政府开放网络实践平台数据、

经济和社会发展数据、教育数据、图书馆文献数据、文化数据和商业数据等，培养形式包括线上线下讲座、参加项目和赛事等实践、软件达人和数据交互社区等，拓展了数据挖掘能力的素养的信息源。

(二) 教学内容设计是以数据挖掘能力培育生命周期为基础，涵盖数据发现与收集、数据清洗与处理、数据挖掘与可视化分析、数据管理与存储和数据共享等系统化过程，有别于原来传统的培育模式，具有创新性。

(三) 设计多元化数据挖掘的实践平台构建，通过把理论知识、统计软件使用以及数据挖掘技术在生产生活中的灵活运用能力进行深度融合的教育教学新途径，更加客观地、高效地孕育和培养数据挖掘的能力目标构成。

基金项目

项目名称：《基于需求驱动的数据挖掘能力培育实践平台的构建》，项目编号：202102647013 教育部 2021 年第二批产学合作协同育人项目。

参考文献

[1] 曾垂省, 舒坤贤, 梁亦龙, 解增言, 谢永芳, 王允, 胡波. 生物信息学专业之数据挖掘教学实践与思考[J]. 广州化工, 2014, 42(7): 190-192.

[2] 汪一百. 网络资源驱动型的《数据挖掘》课程教改分析[J]. 信息与电脑(理论版), 2016(14): 247-248.

[3] 蔡莉, 鲁喜凤. 转型经济下资源驱动型与机会驱动型企业创业行为研究——基于机会与资源的整合视角[J]. 中山大学学报(社会科学版), 2016, 56(3): 172-182. https://doi.org/10.13471/j.cnki.jsysusse.2016.03.018

[4] 赵晓凡. 公安高等院校《数据挖掘》课程教改研究[J]. 计算机教育, 2018(1): 39-42. https://doi.org/10.16512/j.cnki.jsjjy.2018.01.011

[5] 黄发良, 钟世华, 何万莉. 基于 CDIO 理念的《数据挖掘》课程教学探索[J]. 南宁师范大学学报(自然科学版), 2021, 38(2): 191-196. https://doi.org/10.16601/j.cnki.issn2096-7330.2021.02.028

[6] 李姗姗, 李忠. 就业需求驱动下的本科院校《数据挖掘》课程内容体系探讨[J]. 计算机时代, 2015(2): 60-61+64. https://doi.org/10.16644/j.cnki.cn33-1094/tp.2015.02.012