数据挖掘课程案例教学研究

周森鑫,盛鹏飞,王夫芹
(安徽财经大学 管理科学与工程学院,安徽 蚌埠 233030)

摘 要:数据挖掘是一门交叉性学科,是情报学专业的重要课程之一。它主要介绍数据挖掘的基本概念、原理、方法和技术,涉及多个学科和算法因而教学难度较大。由于数据挖掘学科交叉性强涉及的挖掘方法和相关算法多并繁杂,造成学生对数据挖掘的整体工作流程缺乏了解形成"不识庐山真面目只缘身在此山中"现象。文中以时间序列服装销售额预测挖掘项目为教学案例让学生首先掌握数据挖掘的标准流程,重点讲解用到的相关挖掘方法和算法及其在实际挖掘环境中的开发方法,达到"会当临绝顶一览众山小"的教学目标。通过教学实践教学效果良好。

关键词:案例教学;数据挖掘;时间序列;SPSS;Clementine12.0
中图分类号:TP399 **文献标识码**:A **文章编号**:1673-629X(2012)11-0183-04

Research on Case Teaching Method for Data Mining Course

ZHOU Sen-xin, SHENG Peng-fei, WANG Fu-qin
(Management Science and Engineering School of Anhui University of Finance & Economics, Bengbu 233030, China)

Abstract: Data mining course is a interdiscipline and one of the most important courses for intelligence science. It mainly introduces the basic concept, the principle, method and technology of data mining, involving multiple disciplines knowledge and algorithm. Being lack of understanding on the overall engineering process of data mining for the students, thus it is very difficult to handle. It presents a case teaching method, integrating related theory and technology in teaching case, focusing on explaining the use of methods and algorithms for mining and completing them in SPSS Clementine12.0 environment. Through teaching practice, the effect of the case teaching is better.

Key words: case teaching; data mining; time series; SPSS_Clementine12.0

0 引 言

数据挖掘(Data Mining)是从大量的、不完全的、有噪声的、模糊的、随机的实际应用数据中,提取隐含在其中的、事先不知道的、但又是潜在有用的信息和知识的过程,它是一种深层次的数据分析方法。近年来,数据挖掘引起了信息产业界的极大关注,其主要原因信息技术的调整发展带来了爆炸式增长的数据量,可以广泛使用,并且迫切需要将这些数据转换成有用的信息和知识。获取的信息和知识可以广泛用于各种应用,包括商务智能、生产控制、市场分析、工程设计和科学探索等。作为一门交叉性学科,数据挖掘是情报学专业重要的工具之一。数据挖掘的理论来源有两个方面:(1) 来自统计学的抽样、估计和假设检验;(2) 人工智能、模式识别和机器学习的搜索算法、建模技术和学习理论。数据挖掘也迅速地接纳了来自其他领域的思想,这些领域包括最优化、进化计算、信息论、信号处理、可视化和信息检索。同时,数据挖掘还需要得到其它一些学科理论上的支撑和技术上的支持,特别地需要数据库系统提供有效的存储、索引和查询处理支持。高性能计算的技术是处理海量数据集的基础,分布式技术处理海量和分布式数据至关重要。数据挖掘是通过分析每个数据,从大量数据中寻找其规律的技术,主要有数据准备、规律寻找和规律表示三个步骤:

一、数据准备是从相关的数据源中选取所需的数据并整合成用于数据挖掘的数据集;

二、规律寻找是用某种方法将数据集所含的规律找出来;

三、规律表示是尽可能以用户可理解的方式(如可视化)将找出的规律表示出来[1]。

常规的教学过程是讲解数据挖掘的绪论即基本概念、基本理论和相关的挖掘方法。由于数据挖掘学科交叉性强涉及的挖掘方法和相关算法多并繁杂,教学实践中大多采用讲解主要和精典部分,而其余的部分

收稿日期:2012-04-11;**修回日期**:2012-07-16
基金项目:2009年安徽省高校自然基金重大项目(ZD200905);2010安徽省教育教学研究项目(20100473)
作者简介:周森鑫(1965-),男,副教授,硕士生导师,博士,研究方向为数据挖掘、计算机网络、计算机控制;盛鹏飞(1987-),男,硕士研究生,研究方向为数据挖掘。

由学生自学。这种教学模式造成学生对数据挖掘的整体工作流程缺乏了解造成“不识庐山真面目只缘身在此山中”现象。学生对相关的概念和理论理解不深，掌握不透，遇到实际工程项目和科研课题时感觉无从下手。文中提出以典型教学案例为起点让学生首先掌握数据挖掘的标准流程，重点讲解用到的相关挖掘方法和算法及其在实际挖掘方法中的用法，通过挖掘结果比较强化学生感性认识，达到“会当临绝顶一览众山小”的教学效果。通过教学实践此种教学模式操作性强，教学效果良好[2~6]。

1 SPSS Clementine 数据挖掘系统及其基本思想

案例教学模式首先要让学生掌握实际的工程环境和数据挖掘系统的开发流程。在介绍过数据挖掘的基本概念和基本原理后紧接就详细介绍数据挖掘软件开发系统及其使用方法。推荐使用数据挖掘软件开发系统是第三代数据挖掘系统 SPSS 公司的 Clementine12.0，Clementine 提供了大量的人工智能、统计分析的模型（神经网络、关联分析、聚类分析、因子分析等），基于图形化的界面让学生认识、了解、熟悉数据挖掘的全部过程提供了直观的认识。Clementine 还拥有优良的数据挖掘设计思想，数据挖掘过程的每一步工作也很清晰。数据挖掘的目标是从杂乱无章的数据中发现有价值的规则或模式并用于指导实际应用。一般的数据挖掘项目要经历的流程包括问题的理解、数据的理解、收集和准备、建立挖掘模型、模型的评价以及模型的应用等一系列任务。为了对数据挖掘流程进行系统化的抽象，一些数据挖掘流程的参考标准或模型相继被提出。如 SPSS 提出了 5A 标准，SAS 提出了 SEMMA（sample，explore，modify，model，assess），数据挖掘特别兴趣小组则提出了数据挖掘跨行业标准流程，即 CRISP-DM（Cross-Industry Standard Process for Data Mining）[6~8]。

数据挖掘项目的生命周期由六个阶段组成，这些阶段之间的顺序并不固定，常需要根据任务的结果选择执行哪一个特定的阶段。这六个阶段为：商业理解、数据理解、数据准备、模型建立、模型评估以及模型发布。

（1）商业理解（Business Understanding）：商业理解可能是数据挖掘最重要的阶段，在这一阶段要明确面临的商业问题和数据挖掘想要达到的目的，完成商业问题到数据挖掘问题的定义过程。商业理解包括确定业务对象，评估情况，确定数据挖掘目标以及制定工程计划。

（2）数据理解（Data Understanding）：数据提供了数据挖掘的“原材料”。此阶段用于了解数据源以及这些数据的特征，具体任务包括收集初始数据、描述数据、探索数据和验证数据质量。

（3）数据准备（Data Preparation）：对数据源进行分类之后，需要准备数据，以便进行挖掘。数据准备阶段包含了诸多需反复进行的任务，主要是数据预处理，包括数据选择、清理、构建、集成以及格式化数据等。不同的建模工具往往对数据表或记录属性有特定的要求，因此建模前数据的清洗和转换非常重要。

（4）建立模型（Modeling）：模型的建立是数据挖掘的核心部分，在此阶段将使用精巧复杂的分析方法从数据中提取信息，包括建模技术选择、模型的设计与参数校准以及模型的构建与测试等。对于同一类型的数据挖掘问题，可以尝试使用多种方法以比较其各自结果，便于选定最优的模型。

（5）模型评估（Evaluation）：从数据分析的角度，对建立的一个或多个模型及其结果进行对比验证，准确度验证等方法对模型进行详细的评估，以确定模型的价值。此阶段的要素包括评估数据挖掘结果，查看数据挖掘过程，以及确定后续步骤。

（6）模型发布（Deployment）：建立和检验模型并不是数据挖掘的目的，只有把模型发布到相关决策者手中，才能使得通过数据挖掘提高企业决策的科学性，因此建立好适合的模型之后，需要通过发布工具将模型嵌入到用户的应用系统中或者提供完整的报告和测试结果。模型发布后并不意味着一个数据挖掘项目的结束，数据挖掘系统与业务系统间存在着作用与反作用的交互关系，随着时间的推移和数据的变化，模型中的很多关键参数需要及时调整，才能保证挖掘结果的质量并延长其有效的生命周期。Clementine12.0 是以 CRISP-DM 标准流程为主线指导开发者轻松进行数据挖掘项目开发[9,10]。

2 时间序列预测销售额教学案例及实现

在学生掌握 SPSS 公司的 Clementine12.0 的标准数据挖掘流程之后，教学重点是 Clementine 的基本操作方法和技巧，教学方法主要采用演示和让学生实际操作，为案例的讲解作铺垫。教学案例的讲解是最为关键的教学环节，因为通过案例的讲解让学生强化数据挖掘的标准流程并掌握 Clementine 的基本操作方法和技巧并能实际感受挖掘结果。这种方法的好处是先回避较为难理解的相关理论，由实践激发学生的理论学习兴趣，然后讲解相关理论并引导学生自己钻研和深入学习相关理论。例如文中的时间序列服装销售额预测案例它涉及时间序列等相关理论，如果首先讲时间序列理论可能学生由于对相关的数学理论的畏惧从

而失去学习兴趣,影响教学效果。下面针对这个教学案例进行详细介绍,当然在教学过程中根据教学要求和教学目的可选其它的教学案例。建议在选择教学案例时最好选择 Clementine12.0 系统自身演示项目,这些项目提供了数据文件和流文件可使学生自我指导和学习。文中案例流文件的具体位置在\demos\classification_module\catalog_forecast.str,数据文件的具体位置在 \demos\catalog_seasfac.sav。在教学演示开始让学生思考两个问题:销售序列是否有总体趋势,如果有这个趋势会持续还是会随着时间衰减;这个序列是否展现出季节性,如果是季节波动性是随着时间增强还是保持总体不变。

教学演示的关键步骤如下:

1)拖入 SPSS 源节点,选择需要建模的数据文件。把 men 字段的方向置为输出,其余字段的方向置为无。

2)添加一个时间区间节点至 SPSS 源节点。双击时间区间节点,选择时间区间为月,选择从数据构建 ,选择字段为 date。

3)添加一个时间散点图至时间区间节点,增加 men 字段到序列当中,去掉标准化。

4)点击执行可得序列如图 1 所示。

这是销售历史数据时间序列图。因为我们预测因此必须将历史数据中的规律分析清楚。分析该图可知时间序列图显示大体是上升趋势,上升趋势看起来是持续的,展示了其线性趋势;销售额有明显的季节差异,一年中 12 月销售额最高。季节波动性随着时间的推移而增强,因此考虑乘法算法。确认了序列图特征之后,可以开始试着建立模型了。构建一个最优的指数平滑模型包括确定模型类型:模型是否需要包含趋势,季节性或者两者兼顾;然后为模型设置最佳参数。从销售额序列图可以看出,销售额既有线性趋势成分,也有乘数季节性成分,应该用 Winters 模型。重新打开时间序列节点,在建模标签,确认指数平滑依然被选中,点击标准,选择 Winters 乘数法点击确定并点击执行以生成新模型;连接新模型至时间分区节点,连接时间散点图至新模型节点执行可预测如图 2 所示[11]。

Winters 模型既反映了数据的趋势性又反映了数据的季节性。数据集包含十年以及发生在每年十二月的季节性峰值。这十个峰值预测结果跟实际数据相当吻合。然而,预测结果依然过于强调指数平滑的限制。仔细观察上升下降尖锐的部分,该模型还有一些显著的结构没有考虑到。因此要建一个更复杂模型,可以考虑用 ARIMA(自回归求和移动平均模式 autoregressive integrated moving average)过程 。ARIMA 过程能为时间序列创建细微调整的自回归求和移动平均模型。相比指数平滑算法,对于趋势和季节 ARIMA 提供了更为灵活的建模方法。并且允许在模型中增加预测变量的收益。目前为止,考察的是单个变量销售额月数据并用它来解释销售额的变化。其它变量如邮件数、接受订单电话的数量和客户代表数量对于预测还有制约作用。所以多因子模型要比单因子模型好。为了提高预测精度可使用 ARIMA 过程创建一个多因子模型,观

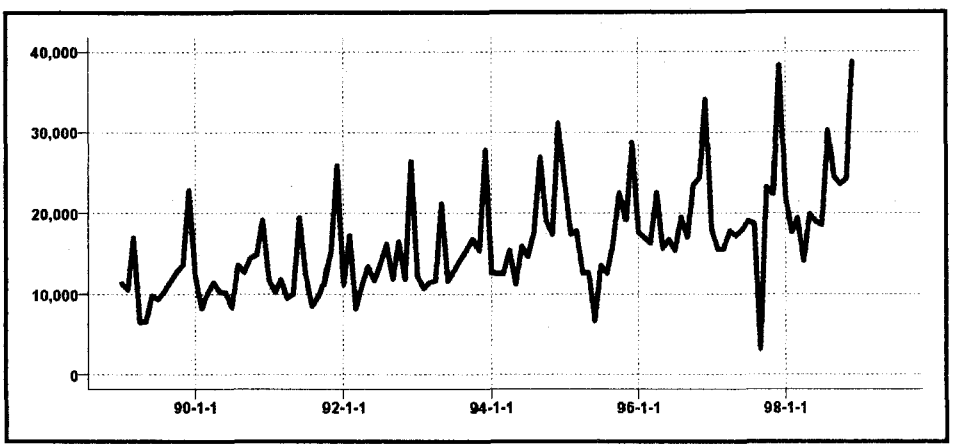

图 1　销售时间序列图

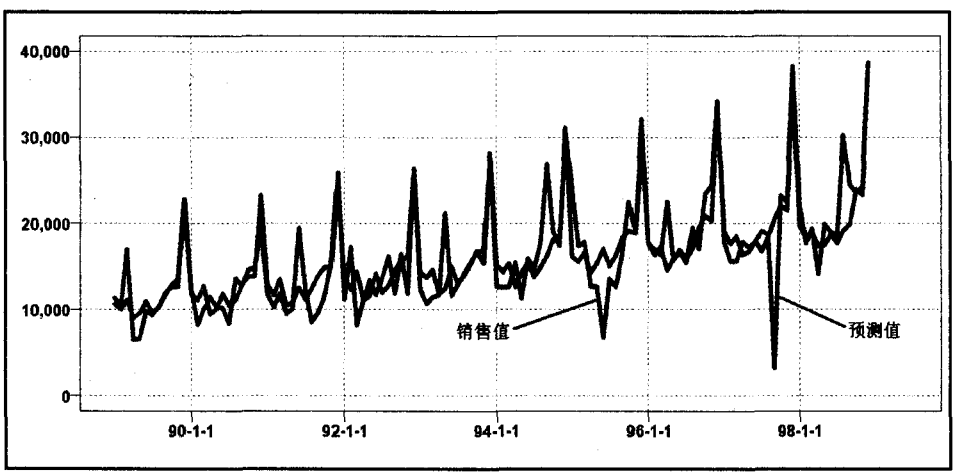

图 2　Winters 模型销售预测图

察是其预测精度的变化。ARIMA 方法能让用户说明自回归的顺序、差分、平均步移。手工决定这些因子的值学生很难把握,因此使用专家模型来选择适用的 ARIMA 模型。将数据集里面的其他一些数据用来建模,包括以下数据:邮寄编目的数量(字段 mail),编目的页数(page),订单热线的数量(phone),打印广告总额(print),客服代表的人数(service)。清除刚才生成的实验模型并打开 SPSS 源节点在类型标签中设置 mail,page,phone,print,service 字段为输入 ,确认 men 为输出,其余的为无点击确定。打开时间序列节点,在模型标签中设置方法为专家建模器,选择仅限于 ARIMA 模型并确认专家建模器考虑季节模型被选中。点击执行生成 ARIMA 模型,将生成的 ARIMA 模型节点连接到时间分区节点并双击打开模型节点。运行生成的模型产生 ARIMA 预测结果如图 3 所示。

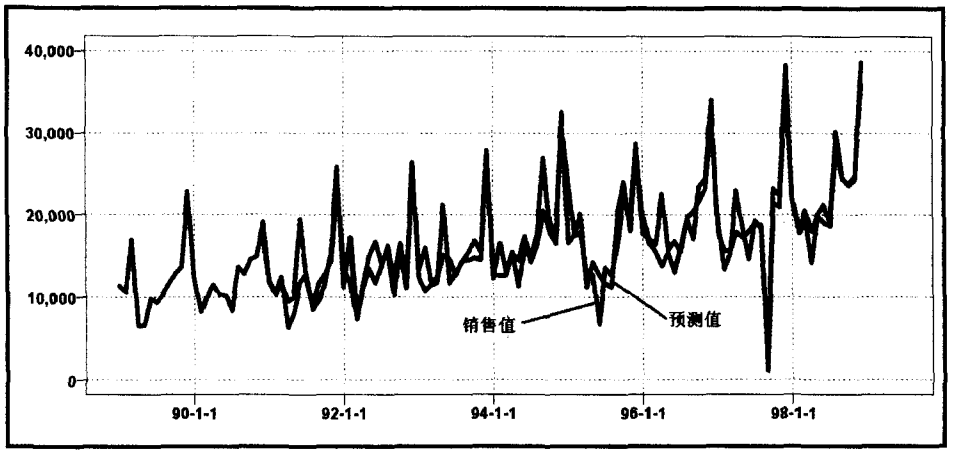

图 3 ARIMA 预测结果图

从图中可发现 1999 年预测精度明显提高:数据在十二月峰值以后回归至正常水平;另外在后半年有个稳定的上升趋势,且明显高于之前的年份。通过案例的演示和讲解学生对时间序列的相关模型有了深刻的感性认识同时观察了不同模型的预测精度差异,接下来再讲解时间序列的指数模型和 ARIMA 模型等相关理论教学效果就会有很大的改进[12~15]。

3 结束语

数据挖掘作为情报学专业的重要课程之一,常规的教学过程是从讲解数据挖掘的绪论即基本概念、基本理论和相关的挖掘挖掘方法开始。由于数据挖掘学科交叉性强涉及的挖掘方法和相关算法多并繁杂,造成学生对数据挖掘的整体工程流程缺乏了解形成"不识庐山真面目只缘身在此山中"现象。文中以时间序列服装销售额预测挖掘项目为教学案例让学生首先掌握数据挖掘的标准流程,重点讲解用到的相关挖掘方法和算法及其在实际挖掘环境中的开发方法,达到"会当临绝顶一览众山小"的教学效果。通过教学实践此种教学模式操作性强教学效果良好。另外可根据不同的教学对象(本科生还是研究生)和课时数组织不同的教学案例,当然教师要对教学案例要有精心准备才能保证教学效果。

参考文献:

[1] Kumar U, Jain V K. Time series models (grey-Markov, grey model with rolling mechanism and singular spectrum analysis) to forecast energy consumption in India[J]. Energy, 2010, 35(4): 1709-1716.

[2] Kayacan E, Ulutas B, Kaynak O. Grey system theory-based models in time series prediction[J]. Expert Systems with Applications, 2010, 37(3): 1784-1789.

[3] Boubaker A, Makram B. Modeling heavy tails and double long memory in North African stock market returns[J]. North African Studies, 2012, 2(3): 195-214.

[4] Claeskens G, Hjort N L. Model Selection and Model Averaging [M]. Cambridge: Cambridge University Press, 2008.

[5] Baheer I A, Hajmeer M. Artificial neural networks: fundamentals, computing, design and application[J]. Microbiological Methods, 2000, 43(1): 3-31.

[6] 汤 琳,何 丰. 隐私保护的数据挖掘方法的研究[J]. 计算机技术与发展,2011,21(4):156-159.

[7] CRISP-DM 1.0 数据挖掘方法论指南[M]. 出版地不祥: CRISP-DM 协会,2000.

[8] 李玲娟,郑少飞. 基于数据处理的数据挖掘隐私保护技术分析[J]. 计算机技术与发展,2011,21(3):94-97.

[9] 段永健. 基于时间序列与支持向量机的信号识别模型及预测[D]. 济南:山东大学,2010.

[10] 谭维敏. 广西电信数据挖掘分析设计与实施[D]. 北京:北京邮电大学,2010.

[11] 刘 学. 电信业务收入预测系统建立及模糊查询应用研究[D]. 大连:大连海事大学,2008.

[12] 艾 玲. 时间序列短期预测的方法和技术[D]. 上海:华东师范大学,2010.

[13] 宫 正,刘晓燕. 时间序列短期趋势信号模型研究[J]. 计算机技术与发展,2011,21(12):105-108.

[14] 朱家元,段宝君,张恒喜. 新型 SVM 对时间序列预测研究[J]. 计算机科学,2003,30(8):124-125.

[15] 易超琴,万建平. 我国电信收入的统计分析[J]. 统计与决策,2005(18):114-115.