

数据挖掘在教育教学中的应用综述

孙中祥¹ 彭湘君¹ 杨玉平² 贺一^{1,3}

(1 重庆师范大学 计算机与信息科学学院, 重庆 400047;

2 重庆电子工程职业学院, 重庆 401331;

3 重庆师范大学 新媒体学院, 重庆 400047)

摘要: 随着教育信息化的不断发展, 基于数据库技术的数据挖掘与教育教学的联系也越来越紧密。通过对国内外数据挖掘在教育领域中的应用研究相关文献进行分析, 从数据挖掘相关技术的角度出发, 总结并归纳了各自在该领域中的应用和研究现状。最后提出了数据挖掘在该领域研究中存在的一些问题与难题以及发展前景。

关键词: 数据挖掘; 教育教学; 发展研究

中图分类号: TP392

文献标识码: A

文章编号: 2095-2163(2012)01-0078-04

Summarization on the Data Mining Application in Education Teaching

SUN Zhongxiang¹, PENG Xiangjun¹, YANG Yuping², HE Yi^{1,3}

(1 College of Computer and Information Science, Chongqing Normal University, Chongqing 400047, China;

2 Chongqing College of Electronic Engineering, Chongqing 401331, China;

3 New Media Academy, Chongqing Normal University, Chongqing 400047, China)

Abstract: With the continuous development of educational informationization, the contact of data mining based on database technology and education teaching becomes more and more close. By analyzing the related articles which is about the applied research of data mining in the educational field at home and abroad, this paper summarizes and generalizes the application and the research status from the perspective of data mining related technologies. Finally it puts forward some problems and challenges, as well as the development prospect of data mining in the educational field.

Key words: Data Mining; Education Teaching; Development Research

0 引言

数据挖掘是一门由多学科交叉融合而形成的新兴学科, 集成了众多学科中成熟的工具和技术, 包括数据库技术、统计学、机器学习、模式识别、人工智能和神经网络等等。随着计算机应用的普及和数据库技术的不断发展, 数据挖掘已经被广泛应用于银行、电信、保险、交通和零售等商业领域。在这些商业领域, 数据挖掘已经取得了众所周知的成绩, 已经能解决诸如数据库营销(Database Marketing)、客户群体划分(Customer Segmentation & Classification)、背景分析(Profile Analysis)、交叉销售(Cross selling)等市场分析行为, 以及客户流失性分析(Churn Analysis)、客户信用记分(Credit Scoring)、欺诈发现(Fraud Detection)等商业问题^[1]。

随着计算机的不断普及和多媒体技术的快速发展, 教育教学模式早已不再只是简单的传统课堂教学方式, 基于网络的课程教育和远程教育已渐次为大家所了解和接受。因而, 各式各样的教育系统和教育信息数据库中积累了大量的数据, 如果将数据挖掘应用于教育教学领域, 就能从这些数据中挖掘出大量有价值的规律来指导和发展教育, 从

而促进教育教学决策的科学化。

本文主要从以下几个部分展开。首先简要地介绍了教育教学中的数据挖掘相关技术, 然后回顾了国内外教育教学发展情况以及教育教学数据的来源并阐述了数据挖掘怎样运用于其中, 之后按照数据挖掘的相关技术, 分组描述了数据挖掘在教育领域的应用情况和研究现状, 最后给出了一些总结与展望。

1 教育教学中的数据挖掘相关技术

数据挖掘(Data Mining), 也可被称作是基于数据库的知识发现(KDD: Knowledge Discovery in Databases), 是指从存放在数据库、数据仓库或其他信息库中的、大量的数据中获取有效的、新颖的、潜在有用的、最终可理解的模式的非平凡过程^[2]。数据挖掘是人工智能与数据库技术相结合的产物, 其中许多方法来源于机器学习。因此, 数据挖掘技术主要有关联规则、序列模式、分类/回归分析、决策树、聚类、神经网络、Web挖掘等等^[3]。下面主要对在教育教学领域中适用的几种数据挖掘技术作简要阐述。

(1) 关联规则。基于数据库的知识发现(KDD)研究领域, 关联规则数据挖掘算法尤为重要。关联规则反映一个

收稿日期: 2011-12-20

基金项目: 本文受到重庆市科委软课题(cstc2011cx-rkx A0341)和重庆师范大学教改课题共同资助。

作者简介: 孙中祥(1988-), 男, 湖北荆州人, 硕士研究生, 主要研究方向: 商务智能与数据挖掘;

彭湘君(1980-), 女, 贵州遵义人, 硕士, 讲师, 主要研究方向: 电子金融与商务智能;

杨玉平(1984-), 男, 重庆人, 硕士, 助理实验师, 主要研究方向: 数字图像处理与模式识别;

通信作者: 贺一(1970-), 男, 四川达县人, 博士, 教授, 硕士生导师, 主要研究方向: 数据挖掘、商务智能、数字媒体技术与创意。

事物与其他事物之间的相互依存性和关联性。如果两个或者多个事物之间存在一定的关联关系,那么,其中一个事物就能通过关联规则由其他事物推测得到。

(2) 分类^[4]。分类可根据某个分类器将数据对象划分到给定的几个类别中的某一类中,一般有两个步骤:构造分类器与利用所获得的分类器对数据进行分类。分类是一种有指导的学习过程,样本训练集是已知类别的,而且类别个数是确定的。分类模型可用于预测,就是利用学习所获得的模型对未知类别的数据对象进行类别判定。

(3) 聚类。聚类是将数据集划分为若干组的过程,并使得同一个组内的数据对象具有较高的相似性,而不同组中的数据对象具有较低的相似性。与分类不同的是,聚类是一种无指导的学习过程,事先不知道样本的类别,也不知道类别的个数。

(4) Web 挖掘。是指利用数据挖掘的原则和思想,将传统的挖掘方法应用到 Web 信息上进行挖掘,从 Web 内容资源、链接结构及用户 Web 使用记录中寻找潜在有用的、用户感兴趣的信息或知识的过程。其所处理的对象包括静态网页、Web 数据库、Web 结构、用户使用记录等信息。通过对这些信息的挖掘,可以得到仅通过文字检索所不能得到的信息。

2 数据挖掘技术在教育教学中的应用

2.1 关联规则在教育教学中的应用

关联规则是使用频率最多,范围最广的数据挖掘技术之一,也可以被广泛地应用于国内外教育教学的决策分析中。运用关联规则的方法可以判定哪一种教学方法适合某类学生或某门课程,从而方便教师进行教学方法的选择,使得分层教学能够在实践中得到更进一步的应用和实施。利用数据挖掘中的关联分析,寻找师生各种行为活动之间的内在联系,从而干预和引导相关的师生行为。通过学校教学数据库中存放的有关课程的历史数据,利用关联规则的挖掘算法,找出影响学生成绩的课程方面的原因,以便科学、合理地进行课程设置^[1]。从大量的教学评价数据中,可挖掘出教学效果与哪些因素有关,从而为教学部门提供决策支持信息^[5]。在远程网络教育方面,关联规则通过各内容之间的关联度,则可挖掘出哪些内容学生将会一起访问。此外,关联规则还能根据学生的借书特点来摆放书籍,方便学生借阅等等。

目前,国内外基于关联规则的数据挖掘在教育教学领域中的研究已经取得一定的成果。霍树勋等^[5]利用关联规则挖掘理论对高校中积累的大量教学评价数据进行挖掘与研究,并通过案例分析,生成了一些有效的关联规则。这些挖掘出的关联规则反映了课堂教学效果与授课教师的职称、教龄、学历、教学手段以及授课对象之间的关系,从而为教学部门提供决策支持信息。吴海鹏等^[6]使用基于 Apriori 不产生候选的多维关联规则挖掘算法,提出了一种高校毕业生就业管理的决策支持系统,为学生培养和就业工作提供了科学的决策。Buldu 和 Uçgün^[7]利用 Apriori 算法对伊斯坦

布尔埃尤普商务 i.m.k.b 职业高中(Istanbul Eyup I.M.K.B. Vocational Commerce High School) 中的学生数据进行挖掘,形成了一些规则,并得到了学生成绩不理想的各门课程之间的关系。García 等^[8]提出了一种基于关联规则的合作教育数据挖掘工具,可以不间断地改善网络课程,而且有助于那些拥有相似课程资料的老师共同分享挖掘到的有用信息。Markellou, Mousourouli, Spiros 和 Tsakalidis^[9]利用 Apriori 算法提出了一个基于本体的框架,这个框架可以决定哪些学习资料更适合推荐给某一类用户。

2.2 聚类在教育教学中的应用

“物以类聚,人以群分”,在自然科学和社会科学中,存在着大量的分类问题。通过聚类,学生组成协作学习小组,即组成具有共同特点、共同学习兴趣的学生群组。这样将便于教师采用不同的教学策略,实施个性化教学与指导。根据读者的借阅信息,采用聚类算法对借阅学生进行分类,可以实现图书馆的个性化服务,更好地激发学生的学习兴趣^[10]。另外,聚类也能用于对 Web 上的文档进行分类,以发现隐藏的有用信息,使用户更好地进行网络学习。传统的聚类算法由于实际应用中数据的复杂性往往失效,因而高维聚类分析已经成为聚类分析的一个研究方向。尽管聚类的应用不如关联规则如此广泛,但在教育教学领域的研究却已呈现蓬勃发展态势,尤其是在网络远程教育方面。

刘美玲等^[4]提出了一种基于 K-Means 算法的成绩聚类分析方法,通过对学生实际成绩进行实例分析,得到一些有用信息,教育者通过这些信息可以实施个性化的教学与指导。吕赛鹤等^[11]基于聚类分析原理,以云南师范大学 26 位计算机公共课教师的教学情况为样本,对教学评价中的十项指标进行了聚类分析,为今后教师应该在哪些方面予以改进和努力提供了有益的参考。Chen 等^[12]综合灰关联理论(GRA)、K-means 聚类算法、模糊关联规则和模糊推理四种计算智能理论,提出了基于数据挖掘的学习绩效评估系统。通过该系统,老师可以知晓影响网络教育环境中学习者学习成绩的根本原因。Talavera 和 Gaudioso^[13]利用聚类算法挖掘学生信息以发现可反映用户行为的有效模式,并基于协同管理方案提出了模型,能够在非结构化的合作空间中描述相似行为群体的特征。

2.3 分类在教育教学中的应用

分类算法中应用得最多的是决策树算法。决策树是以实例为基础的归纳学习算法。决策树主要用于对离散数据进行分类,在教学方面则能够用来对学生课程的选修、毕业课题的选择、就业等方面进行分析指导^[14]。利用 ID3 分类算法对学生的历史综合测评数据进行分析,从而构造综合素质评分的分类器,以改进综合测评的方法,促进学生德、智、体各方面的全面发展。在远程教育中,分类可用来锁定那些学习热情不高者并找出补救措施以减少中途退学率,还可预测哪些课程将会取得成功。决策树作为分类算法中的一种,也是使用频率最多、范围最广的数据挖掘技术之一,广泛地应用于各个领域之中。

唐一,马征^[15]以某高校就业信息系统为基础,对其中积

累的海量数据运用决策树 ID3 改进算法进行挖掘并抽取规则知识,从而为学校管理者在提高就业率、提高就业层次、改进现行培养机制等方面提供决策支持。Hsia 等^[16]运用决策树算法来研究分析台湾某高校的继续推广教育中心所开展的课程中哪些课程最受用户喜爱,以吻合各类用户需求及满足为地方培养更多、更好人才的需要。Wang 等^[17]通过对学生个人信息使用决策树算法研究而提出了一种自适应分析系统,可帮助用户优化学习顺序,为每一个特定的教学内容制定出最为适合的学习顺序。Hamalainen 等^[18]提出了一个贝叶斯网络模型来描述学生的学习过程,可以根据学生的技能和其他特征对学生进行分类并进行富含针对性的引导。

2.4 Web 挖掘在教育教学中的应用

Web 挖掘就是使用挖掘技术在 WWW 数据中发现潜在的、有用的模式或信息,这也是目前数据挖掘研究的热点。在网络环境下,针对用户的使用记录和基本信息,采用 Web 挖掘技术分析用户的使用习惯和个人爱好,可为用户提供与众不同的个性化服务。Web 挖掘通过挖掘用户的行为记录和反馈情况为网络教育站点的设计者提供改进的依据,比如应如何组织页面链接结构等。Web 挖掘也可以通过用户的拥塞记录发现站点的性能瓶颈,进而提示网络教育站点的管理者需要将 Web 缓存策略、网络传输策略、流量负载均衡机制和数据的分配策略加以改进^[19]。此外,Web 挖掘还能用来分析用户访问 Web 站点的行为特点,例如,如果有哪些页面用户只要访问了其中一页,就可以推断出该用户是否还要访问其他的页面等。

随着互联网技术的发展和运用,基于 Internet 技术的网络教育越来越普及,因而 Web 挖掘也较广泛地应用于教育教学领域的研究中。如梁燕红^[20]利用 Web 日志挖掘技术进行实例分析,找出访问频度较高的知识点网页,确定学生的兴趣点及知识点中的难点所在。在此基础上,帮助教育者调整教学策略,改善网络教学效果。卢朝晖,刘家宁^[21]提出了一个基于 Web 日志挖掘技术的教学反馈系统,试图揭示学生用户日志中的内在联系,发现隐藏的信息与规律,方便教师了解学生对所学知识的掌握情况与需求,动态地调整教学策略。Romero 等^[22]利用 Web 挖掘技术提出了一个先进的个性化系统架构,并开发出了一种特定的 Web 挖掘工具,将其和一种推荐搜索引擎一同整合进了 AHA! 网络教育系统。该系统能够提供非常个性化的服务,引导学生点击最合适的链接,方便学生学习。Tane, Schmitz, 和 Stumme^[23]提出了一种基于本体论的工具,以便能充分利用 Web 上的有效资源。该工具可根据 Web 文档内容的相似性和主题,利用文本挖掘和文本聚类技术对这些文档进行分组,从而方便用户更好地查询和组织网络资源,提高学习效率。

3 数据挖掘技术在教育教学领域应用中存在的问题

虽然数据挖掘技术在教育教学领域的应用较为广泛,也出现了一些新的尝试,但还只算是一个比较年轻的研究领域,迄今为止,仍存在着一些问题。主要包含以下几个方

面:

(1) 数据质量。许多数据库很可能是动态的、错误的且不完美的。因此在恰当使用数据挖掘技术进行挖掘时,要进行噪声数据处理。

(2) 信息人员的技能。要求信息人员具备较高的计算机水平,懂得数据挖掘和数据仓库方面的知识,能够提出决策支持方面的需求。同时要求其能够分析和试验各种数据挖掘技术,以发现大量的潜在模式,然后以可以理解的形式进行发布。

(3) 性能和成本。为了满足许多数据挖掘系统的计算要求,需要在硬件、操作系统软件和数据库系统等方面采用并行技术。这些资源大大增加了成本,不利于数据挖掘在教育领域普及。

4 数据挖掘技术在教育教学领域应用的展望

相比于数据挖掘在诸如商业等其他领域的应用,数据挖掘技术在教育教学领域的应用起步比较晚,但也已取得了一定的研究成果。但是,要想达到数据挖掘在商业等其他领域的成就,数据挖掘在教育领域的研究还有许多工作要做。这些研究包括^[24]:

(1) 挖掘工具简单实用,面向公众。目前的大多数挖掘工具对教育者和管理者来说都过于复杂,只有信息人员才能使用自如。因此,进行类似于 windows 操作系统那种不需要过多相关专业知识的图像界面挖掘工具的研究就显得很有必要。

(2) 挖掘工具标准化。在网络教育方面,当前的挖掘工具都是针对某一具体的网络教育系统而言的,还没有适用于所有系统的通用挖掘工具出现。

(3) 特定的数据挖掘技术。把教育教学领域的知识与数据挖掘技术有效地结合起来而形成特定的数据挖掘技术,专门研究教育教学领域的问题。特定的教育数据挖掘技术能为教育者提供更加科学的决策,而不像传统的数据挖掘技术需要不断调整参数与算法以满足教育研究的需求。

此外,云计算的并行计算能力和海量存储能力在为教育信息化带来便利和优惠的时候,也使得基于教育数据的传统数据挖掘变得更加困难与复杂。数据复杂度与系统计算能力形成鲜明对比,因此,基于云计算的并行数据挖掘应用于教育教学领域已成为发展的必然。在未来,随着教育数据的急剧增加,基于云计算平台的并行数据挖掘在教育领域的应用会越来越受人们关注。

参考文献:

- [1] 杨永斌. 数据挖掘技术在教育中的应用研究[J]. 计算机科学, 2006(12):284-286.
- [2] 李雄飞,董元方,李军,等. 数据挖掘与知识发现[M]. 第2版. 北京:高等教育出版社,2010:2-5.
- [3] SOMAN K P,等著. 数据挖掘基础教程[M]. 范明,牛常勇,译. 北京:机械工业出版社,2009:1-25.
- [4] 刘美玲,李熹,李永胜. 数据挖掘技术在高校教学与管理中的应用[J]. 计算机工程与设计,2010(5):1130-1133.
- [5] 霍树勋,尹作友. 关联规则挖掘理论在教学评价中的应用[J].

软件导刊,2009(2):11-12.

- [6] 吴海鹏,马吉权. 决策支持系统在高校毕业生就业管理中的应用研究[J]. 黑龙江高教研究,2008(6):48-50.
- [7] BULDU A, ÜÇGÜN K. Data mining application on students' data. *Procedia Soc Behav Sci* 2010:5251-5259.
- [8] GARCÍA E, ROMERO C, VENTURA S, de CASTRO C. A collaborative educational association rule mining tool, *Internet and Higher Education*. doi:10.1016/j.iheduc.2010.07.006.
- [9] MARKELLOU P, MOUSOUROULI I, SPIROS S, et al. Using semantic web mining technologies for personalized e-learning experiences[C]// *Proceedings of the web-based education*,2005:46-1-826.
- [10] 茹蓓,赵芳. 聚类算法在图书馆中的应用[J]. 新乡学院学报(自然科学版),2011(1):41-43.
- [11] 吕赛鹤,李志平. 聚类分析方法在高校教学评价中的应用[J]. 现代教育技术,2009(19):34-36.
- [12] CHEN C M, CHEN Y Y, LIU C Y. Learning performance assessment approach using web-based learning portfolios for e-learning systems. *IEEE Trans Syst Man Cybern C Appl Rev*, 2007,37(6).
- [13] TALAVERA L, GAUDIOSO E. Mining student data to characterize similar behavior groups in unstructured collaboration spaces[C]// *Workshop on artificial intelligence in CSCL. 16th European conference on artificial intelligence*,2004:17-23.
- [14] 方芳,叶春明,曾广军. 数据挖掘技术在教学中的应用研究[J]. 电脑知识与技术,2010(1):143-144,147.
- [15] 唐一,马征. 改进的ID3算法在高校就业系统中的应用分析[J]. 科技情报开发与经济,2011(4):155-157.
- [16] HSIA Taichang, SHIE Anjin, CHEN Lichen. Course planning of extension education to meet market demand by using data mining techniques - an example of Chinkuo technology university in Taiwan[J]. *Expert Systems with Applications*, 2008,34:596-602.
- [17] WANG Y H, TSENG M H, LIAO H C. Data mining for adaptive learning sequence in English language instruction. *Expert Syst Appl*, 2009,36:7681-7686.
- [18] HAMALAINEN W, SUHONEN J, SUTINEN E et al. Data mining personalizing distance education courses[C]// *World conference on open learning and distance education*, Hong Kong, 2004.
- [19] 王利. 数据挖掘技术在高校管理中的应用[J]. 福建电脑,2005(6):248-49.
- [20] 梁燕红. Web挖掘技术在网络教学中的应用[J]. 计算机网络,2011:656-657.
- [21] 卢朝晖,刘家宁. Web日志挖掘技术在网络教学平台中的应用研究[J]. 制造业自动化,2009(11):184-185,188.
- [22] ROMERO C, VENTURA S, ZAFRA A, et al. Applying Web usage mining for personalizing hyperlinks in Web-based adaptive educational systems. *Comput Educ*, 2009,53:828-840.
- [23] TANE J, SCHMITZ C, STUMME G. Semantic resource management for the web: Ane-learning application[C]// *Proceedings of the WWW conference*, New York, USA, 2004:1-10.
- [24] ROMERO C, VENTURA S. Educational data mining: A survey from 1995 to 2005[J]. *Expert Systems with Applications*, 2007,(33):135-146.
- [25] KOTISIANIS S B. Use of machine learning techniques for educational proposes: a decision support system for forecasting students' grades[J]. *Springer Science+Business*, 2011.
- [26] 常桐善. 数据挖掘技术在美国院校研究中的应用[J]. 复旦教育论坛,2009,7(2):72-29.
- [27] 顾小清,黄景碧,朱元琨,等. 让数据说话:决策支持系统在教育中的应用[J]. 开放教育研究,2010,16(5):99-106.
- [28] SHAO Yuxiang, CHEN Qing, YIN Weiming. The application of improved decision tree algorithm in data mining of employment rate: evidence from China[J]. *2009 First International Workshop on Database Technology and Applications*, 2009(72):202-205.
- [29] 李瑞林. 数据挖掘技术在教学过程中的应用[J]. 制造自动化,2010,32(9):153-154,211.
- [30] MARTIN S, DIAZ G, SANCRISTOBAL E, et al. Technology trends in education: Seven years of forecasts and convergence[J]. *Computers & Education*, 2011,57:1893-1906.
- [31] 李汪丽,蒋玲. 数据挖掘在中国教育领域的应用研究综述[J]. 湖北第二师范学院学报,2010,27(2):55-58.
- [32] GURULER H, ISTANBULLU A, KARAHASAN M. A new student performance analysing system using knowledge discovery in higher educational databases[J]. *Computers & Education*, 2010,55:247-254.
- [33] PAN Qingxian, QU Linjie, LOU Lanfang. Data mining and application of teaching evaluation based on association rules[J]. *Proceedings of 2009 4th International Conference on Computer Science & Education*, 2009(9):1404-1407.
- [34] 孙柏祥. 云计算—高校教育信息化建设和发展的新模式[J]. 中国电化教育,2010(5):123-125.
- [35] 李凯,常征. 基于云计算的并行数据挖掘系统设计与实现[J]. 微计算信息,2011(6):121-123.