

# Moodle 平台数据挖掘方法设计与实现

张国荣

(广州美术学院 艺术与人文学院, 广东 广州 510260)

**摘要:**教育数据挖掘是一个新兴的研究方向。如何把存储在教育软件系统中的数据转变为有意义的信息,并为教育决策、优化教学过程服务,已成为大多数教育工作者所关注的内容。文中总结了当前教育数据挖掘的研究现状,介绍了一种基于 Excel 的简单数据挖掘方法。该方法利用模糊 C 均值聚类算法,对 Moodle 平台积累的日志数据进行分析,找出有相似学习行为的学生,为学习社区的小组划分和研究学习模式服务。实验表明,该方法能够更准确地对学生进行分类,而且操作更为简单、方便。

**关键词:**教育数据挖掘;聚类;日志挖掘;Moodle

中图分类号:TP301

文献标识码:A

文章编号:1673-629X(2014)05-0231-04

doi:10.3969/j.issn.1673-629X.2014.05.057

## Design and Realization of Data Mining Method in Moodle

ZHANG Guo-rong

(School of Arts and Humanities, Guangzhou Academy of Fine Arts,  
Guangzhou 510260, China)

**Abstract:** Educational Data Mining (EDM) is an emerging research direction. How to use the massive educational data and transfer the data into useful information and knowledge in order to provide the service for educational decision and teaching optimization processing has become an emerging and concerned research domain. It reviews the status on educational data mining, and proposes a simple data mining method based on Excel, which analyzes log data using fuzzy c-means clustering algorithm to identify students who have similar learning behavior and evolve hidden patterns in the Moodle. The experiment demonstrates that the method can more accurately classify the students, and the operation is more simple and convenient.

**Key words:** educational data mining; clustering; log mining; Moodle

## 0 引言

数据挖掘已被成功地应用于商业领域,成为市场营销、市场预测、欺诈防范等工作流程中重要的技术手段。随着教育领域数据收集量的不断增加以及教育决策对量化分析结果的依赖加强,数据挖掘在教育管理中的研究与应用呈显著上升趋势。自2005年起,有多个重要的国际会议开展了以“教育数据挖掘”为主题的研讨会。2007年,研究者在第二届欧洲技术促进学习会议之后组成国际教育数据挖掘工作组,创办在线学术期刊——教育数据挖掘杂志(JEMD),并从2008年开始每年召开“教育数据挖掘国际会议”,2013年7月,第六届教育数据挖掘国际会议将在美国的孟菲斯举办<sup>[1]</sup>。

## 1 教育数据挖掘

根据教育数据挖掘杂志(JEMD)的定义,教育数据挖掘是指从教育数据中挖掘有意义信息的过程,这些信息可以为教师、学生、教育管理者甚至软件开发者和研究者等提供服务<sup>[2]</sup>。对于不同的人群,教育数据挖掘有其特定的价值。对于学习者而言,教育数据挖掘能够向学习者推荐有助于改进他们学习的学习活动、学习资源、学习经验和学习任务。这些建议可以通过分析这些学习者完成的行为以及与之相似的学习者完成的行为来取得。对于教育工作者而言,教育数据挖掘能够向他们提供更多更客观的反馈信息,使他们能够更好地调整和优化教育决策、改进教育过程、完善课程开发,并根据学习者的学习状态来组织教学内容、重构教学计划等<sup>[3]</sup>。

收稿日期:2013-07-10

修回日期:2013-10-16

网络出版时间:2014-02-11

基金项目:广东省教育科研“十二五”规划2012年度研究项目(2012JK190)

作者简介:张国荣(1977-),男,讲师,硕士,CCF会员,研究方向为计算机基础教学、教育数据挖掘。

网络出版地址: <http://www.cnki.net/kcms/detail/61.1450.TP.20140211.1453.022.html>

教育数据挖掘从技术方法角度进行分类,可以划分为:统计分析可视化、聚类(聚类、离群点分析)、预测(决策树、回归分析、时序分析)、关系挖掘(关联规则挖掘、序列模式挖掘、相关分析)和文本挖掘等<sup>[3-5]</sup>。从具体的应用领域角度进行分类,又可以划分为:E-Learning 数据挖掘、E-Management 数据挖掘和 E-Research 数据挖掘<sup>[3]</sup>。当然,对于数据挖掘的具体应用,国内外已有较成熟的综述文献。文献[1]运用文献计量与内容分析法,对国内外公开发表的关于教育数据挖掘的文献进行统计分析,把握其发展脉络及研究现状,探讨研究中的关键内容,并展望该领域未来的研究趋势,为进行教育数据挖掘的研究与实践提供参考文献。文献[6]系统梳理了国内外 E-Learning 数据挖掘的研究进展,并采用格语法分析方法对“E-Learning”的关键要素和过程进行分析,提出以“谁在学、学什么、怎么学、学得如何”这一系列问题为主线,开展 E-Learning 数据挖掘工作,从而获得对 E-Learning 现状的更加完整的认识。文献[7]总结了数据挖掘技术在教育领域的应用与历史,分别介绍了数据挖掘技术在传统教育系统、远程教育系统、智能辅导系统和在线学习系统上的应用,重点描述了如何把主要的数据挖掘技术应用在教育数据上。文献[8]则主要关注学生的流失和个人推荐系统在教育上的应用等主题,并详细分析数据挖掘技术如何应用在课程管理系统的数据库上,指出当前研究的不足和未来研究的方向。

对于成功应用数据挖掘技术的案例,国内文献也有详细介绍。文献[9]以一个实际运行的移动学习网站为研究平台,利用数据挖掘技术对网站客观数据进行分析研究,并就“移动学习使用者特征”、“知识资源个性化推荐技术”和“资源需求趋势预测”这三个比较值得关注的问题进行了实验论证,对移动学习研究者和基于知识服务的网站运营决策者具有一定的启发意义和参考价值。文献[10]借鉴国外相关成果开展 Moodle 平台数据挖掘研究,通过常规统计方法、可视化方法、聚类方法、社会网络分析等方法,对网络学习平台的日志和交互论坛数据进行数据挖掘,揭示出某一网络培训班在线学习过程中师生活动的总体情况,发现学生的模块访问偏好和学习时间偏好,分析得出师生交互网络的结构特点。这些挖掘结果可为评估网络培训开展情况、优化学习支持服务等提供参考。

但是,数据挖掘技术是一种面向应用的复杂技术,应用难度很大。很多数据挖掘工具是针对商业用途开发的,例如 DBMiner、SPSS Clementine、SAS Enterprise Miner、IBM Intelligent Miner 等。这些工具不是专门为教育领域而设计,对很多教育工作者来说设计的过于复杂,不易于使用<sup>[1]</sup>。在有效准确挖掘教育数据的同

时,如何针对教育领域的特点,开发一些直观、易于使用的工具,以帮助教师对教学过程和学生的学习行为进行分析是一个非常有意义的问题。文中将介绍一种基于 Excel 的简单数据挖掘方法,该方法利用模糊 C 均值聚类算法,对 Moodle 平台积累的日志数据进行分析,找出有相似学习行为的学生,为学习社区的小组划分和研究学习模式服务。实验分析表明,该方法能够更准确地对学生进行分类,而且操作更为简单、方便。

## 2 模糊 C 均值聚类算法

聚类是一个将物理或者抽象对象的集合分组成为类似对象组成的多个类或簇的过程<sup>[11]</sup>。用聚类方法划分相似学生群体或个体,可以提供相似或个性化的教学。同时,课程学习过程经常需要形成小组进行学习,根据学生的学习习惯,运用聚类算法可以合理地分配小组成员。文中选用应用广泛的模糊 C 均值聚类算法对 Moodle 平台积累的日志数据进行分析,希望找出有相似学习行为的学生。

模糊 C 均值(FCM)聚类算法是理论最为完善、应用最为广泛的基于目标函数的模糊聚类算法<sup>[12]</sup>。模糊 C 均值聚类算法是用隶属度确定每个数据点属于某个聚类的程度的一种聚类算法。它把  $n$  个向量  $x_i(i = 1, 2, \dots, n)$  分为  $c$  个模糊组,并求每组的聚类中心,使得非相似性指标的目标函数达到最小。FCM 应用模糊划分,使得每个给定数据点用值在 0,1 间的隶属度来确定其属于各个组的程度。与引入模糊划分相适应,隶属矩阵  $U$  允许取值在 0,1 间的元素<sup>[12]</sup>。

定义 1:给定  $p$  维空间  $R^p$  中的数据对象集  $X = \{x_1, \dots, x_n\}$ , 设  $\tilde{C}_1, \dots, \tilde{C}_c$  分别表示  $X$  上的  $c$  个模糊集,  $\chi_{ij}$  表示  $x_j$  属于  $\tilde{C}_i$  的隶属度。若

$$\sum_{i=1}^c \chi_{ij} = 1, 1 \leq j \leq n \quad (1)$$

$$0 < \sum_{j=1}^n \chi_{ij} < n, 1 \leq i \leq c \quad (2)$$

那么集合  $\tilde{C}_1, \dots, \tilde{C}_c$  称为  $X$  的  $c$ -模糊划分

定义 2:对于任意一个数据对象  $x_j$  和一个以  $y_i$  为中心的簇,设  $d_{ij}$  表示  $x_j$  和  $y_i$  间的欧氏距离,即

$$d_{ij} = \|x_j - y_i\|, 1 \leq i \leq c, 1 \leq j \leq n \quad (3)$$

那么,  $x_j$  属于簇  $C_i$  的隶属度定义为

$$\chi_{ij} = \left[ \sum_{k=1}^c \left( \frac{d_{kj}}{d_{ij}} \right)^{2/(\beta-1)} \right]^{-1}, 1 \leq i \leq c, 1 \leq j \leq n \quad (4)$$

其中,若对所有  $1 \leq k \leq c, d_{kj} \neq 0$ , 而  $\beta > 1$  是一个用于聚类过程中模糊度控制的调整参数,若对于某个  $i$  和  $j$ ,  $d_{ij} = 0$ , 那么定义  $\chi_{ij} = 1; \chi_{kj} = 0, k \neq j$ 。

定义 3:模糊聚类的过程中,需要寻找分别以  $y_1, \dots, y_c$  为中心的  $c$  个簇,使目标函数值最小,目标函数定义如下:

$$I_f = \sum_{j=1}^n \sum_{i=1}^c \chi_{ij} \|z_i - x_j\|^2; z_i \in R^n, 1 \leq i \leq c \quad (5)$$

其中,  $z_i (1 \leq i \leq c)$  是未知的聚类中心,FCM 算法使用上次迭代过程中的隶属度值修改  $z_i$ ,并使  $c$  个目标函数的值局部最小,即局部最小化下式的值:

$$I_f(i) = \sum_{j=1}^n \chi_{ij} \|z_i - x_j\|^2, 1 \leq i \leq c \quad (6)$$

聚类的迭代过程中,任何一个模糊聚类  $\tilde{C}_i$  的中心做如下的修改:

$$y_i^{(k+1)} = \frac{\sum_{j=1}^n \chi_{ij}^{(k)} x_j}{\sum_{j=1}^n \chi_{ij}^{(k)}} \quad (7)$$

其中,  $\chi_{ij}^{(k)}$  表示经过第  $k$  次迭代后  $x_j$  属于  $\tilde{C}_i$  的隶属度。

具体算法:

输入:

$n$ :数据对象的个数;

$c$ :结果簇的个数;

$X = \{x_i\}, 1 \leq i \leq n$ :  $p$  维空间中  $R^n$  中的数据集;

$Y_0 = \{y_{i0}\}, 1 \leq i \leq n$ :最初的  $c$  个簇中心;

$N$ :所允许的最大迭代次数;

$\varepsilon$ :簇间距离的阈值;

$\beta$ :模糊度控制参数。

输出:

$Y = \{y_i\}, 1 \leq i \leq c$ :结果簇中心;

$(\chi_{ij}), 1 \leq i \leq c, 1 \leq j \leq n$ :结果隶属度矩阵;

it:迭代次数。

步骤 1

初始化:令  $k = 0, y^{(0)} = y_{i0}, 1 \leq i \leq c$

步骤 2

for  $i = 1$  to  $c$  do

for  $j = 1$  to  $n$  do

$d_{ij}^{(k)} = \|x_j - y_i^{(k)}\|$

end for

end for

步骤 3

for  $i = 1$  to  $c$  do

for  $j = 1$  to  $n$  do

$\chi_{ij} = \left[ \sum_{k=1}^c \left( \frac{d_{ij}}{d_{kj}} \right)^{2/(\beta-1)} \right]^{-1}$

end for

end for

步骤 4

if  $d_{ij}^{(k)} = 0$  (对某个  $l = l_0$ ) then

对所有  $i \neq l_0, \chi_{l_0}^{(k)} = 1, \chi_{ij}^{(k)} = 0$

end if

步骤 5

for  $i = 1$  to  $c$  do

修改簇中心:  $y_i^{(k+1)} = \frac{\sum_{j=1}^n \chi_{ij}^{(k)} x_j}{\sum_{j=1}^n \chi_{ij}^{(k)}}$

end for

步骤 6

if  $\left[ \sum_{i=1}^c \|y_i^{(k+1)} - y_i^{(k)}\|^2 \right]^{1/2} < \varepsilon$  then

$y_i = y_i^{(k+1)}, 1 \leq i \leq c$

$\chi_{ij} = \chi_{ij}^{(k)}, 1 \leq i \leq c, 1 \leq j \leq n$

it =  $k + 1$

for  $i = 1$  to  $c$  do

输出  $y_i, \chi_{ij}$ , it

结束算法

end for

end if

步骤 7

if  $k = N$  then

输出“算法不收敛”

算法结束

else

转到步骤 2

end if

### 3 模糊 C 均值聚类 Moodle 平台上的应用

Moodle 是一个免费的开放源代码的学习管理系统,目前我国教育领域得到了广泛应用,有许多围绕 Moodle 平台的应用研究。然而,很少有人利用 Moodle 平台强大的日志功能开展数据分析研究。文献[10]是国内对 Moodle 平台日志功能开展数据分析的典型,但仅仅是演示如何进行聚类,并没有讨论详细的算法,而且简单的聚类生硬的把学生分组可能是不合适的。文中利用模糊聚类的算法,在 Excel 上利用 VBA 语言开发聚类工具,并对 Moodle 平台的日志数据进行分析挖掘,结果表明,工具能够有效地对数据进行聚类,同时,工具不需要特殊的安装环境,简单易用。

#### 3.1 数据采集与预处理

2012 年开始,在广州美术学院进行“翻转课堂”教学模式实验,并在 Moodle 教学平台上建立了网络课程,积累大量的日志数据,这里以《CSS 网页设计》课程

的同学为研究对象,根据 Moodle 平台 (<http://www.zhangguorong.net>) 积累的学习数据进行模糊聚类研究。

课程从 2012 年 9 月开始,2013 年 1 月结束,共有 79 位同学参加学习,Moodle 平台的日志数据表记录了用户详细的访问信息,包括了每一个用户所访问的平台模块、各种操作以及发生的时间。利用这一日志数据表,可以对该课程学生访问平台模块和各类操作行为情况等统计和聚类分析。

以管理员的身份登录 Moodle 平台,可以选择 Excel 文件格式下载所有日志。对 Excel 表的数据进行处理,统计学生访问各个模块的次数,选择 5 个学生作为样本进行分析。表 1 是学生分别访问作业 (assignment)、课程 (course)、论坛 (forum)、资源 (resource) 这 4 个模块的频次统计样例。

表 1 学生访问学习平台模块的频次统计样例

编号	assignment	course	forum	resource
1001	30	120	233	9
1002	40	55	175	35
1003	19	86	81	25
1004	27	45	95	18
1005	15	50	166	13

3.2 模糊 C 均值 (FCM) 聚类

设定  $c = 2, N = 15, \varepsilon = 0.005, \beta = 2$ , 经过 15 次的迭代运算,结果类中心和隶属度矩阵如表 2 和表 3。

表 2 模糊聚类分析后的类中心

类中心	assignment	course	forum	resource
$C_1$	21.99	55.10	110.44	18.09
$C_2$	27.89	69.50	192.00	13.70

表 3 模糊聚类分析后学生的隶属度

类别	1001	1002	1003	1004	1005
第一类	0.11	0.33	0.92	0.96	0.98
第二类	0.89	0.67	0.08	0.04	0.02

经过模糊聚类分析,把学生分成两类,可以看到,编号 1001 和 1002 的学生属于同一类,他们有类似的访问习惯,而编号 1003、1004 和 1005 的访问习惯更为接近,而且隶属度也显示他们更为明确地属于第一类,对于编号 1002,虽然他可以划分属于第二类,但显然他也有部分与第一类相似。通过模糊聚类分析,在划分小组进行个性化教学时,可以更好地理解学生分别属于哪一类。

3.3 模糊聚类工具

为了能够更简单方便的进行模糊聚类分析,没有选用复杂的商用软件,而是在 Excel 上开发桌面版的

聚类分析工具,利用 Visual Basic for Applications (VBA),设计实现了一种有效、容易使用的聚类工具。工具不需要特殊的安装,只要计算机上安装了微软公司 Microsoft Excel 就可以进行聚类分析,简单易用。

4 结束语

随着大数据时代的到来,教育数据挖掘成为一个新兴的研究方向,越来越受到教育工作者的关注。文中利用模糊 C 均值聚类算法,通过挖掘 Moodle 平台日志,并根据学习者日常在学习平台的访问行为进行模糊聚类计算,这些挖掘得到的信息将有助于改进教师的教学工作。文中的创新之处在于:使用模糊聚类的方法对教育数据进行分析,为学生分组提供更多的信息。同时,提供了一种简便的分析工具,在常用的 Excel 软件上进行开发,使普通的教师也能进行模糊聚类分析。

参考文献:

[1] 李 婷,傅钢善. 国内外教育数据挖掘研究现状及趋势分析[J]. 现代教育技术,2010,20(10):21-25.

[2] Baker R. Journal of educational data mining[J/OL]. 2008. <http://www.educationaldatamining.org/JEDM/>.

[3] 葛道凯,张少刚,魏顺平. 教育数据挖掘:方法与应用[M]. 北京:教育科学出版社,2012.

[4] Romero C, Ventura S. Educational data mining: A survey from 1995 to 2005[J]. Expert Systems with Applications, 2007, 33(1):135-146.

[5] Baker R S J D, Yacef K. The state of educational data mining in 2009: A review and future visions[J]. Journal of Educational Data Mining, 2009, 1(1):3-16.

[6] 葛道凯. E-Learning 数据挖掘:模式与应用[J]. 中国高教研究, 2012(3):8-14.

[7] Sachin R B. A survey and future vision of data mining in educational field[C]//Proc of 2012 second international conference on advanced computing & communication technologies. [s. l.]:[s. n. ], 2012:96-100.

[8] Huebner R A. A survey of educational data-mining research [J/OL]. 2013. <http://www.aabri.com/rhej.html>.

[9] 刘 钢,王敏娟,张 驰,等. 移动学习中的数据挖掘研究[J]. 中国远程教育, 2011(1):31-35.

[10] 魏顺平. Moodle 平台数据挖掘研究-以一门在线培训课程学习过程分析为例[J]. 中国远程教育, 2011(1):24-30.

[11] Han Jiawei, Kamber M. 数据挖掘:概念与技术[M]. 北京:机械工业出版社, 2001.

[12] 刘惟一,李维华,岳 昆. 智能数据分析[M]. 北京:科学出版社, 2007.

# Moodle平台数据挖掘方法设计与实现

作者：[张国荣, ZHANG Guo-rong](#)

作者单位：[广州美术学院 艺术与人文学院, 广东 广州, 510260](#)

刊名：[计算机技术与发展](#) 

英文刊名：[Computer Technology and Development](#)

年, 卷(期): 2014(5)

本文链接: [http://d.wanfangdata.com.cn/Periodical\\_wjfz201405057.aspx](http://d.wanfangdata.com.cn/Periodical_wjfz201405057.aspx)