文章编号:1000-5862(2016)06-0640-04

面向虚拟学习社区的学习行为特征挖掘与分组方法的研究

程 艳[1,2],解建华[1],谭平飞[1],杨志明[1]
(1. 江西师范大学计算机信息工程学院,江西 南昌 330022;2. 同济大学计算机科学与技术博士后流动站,上海 201804)

摘要:虚拟学习社区作为网络教育的一种新兴模式,越来越受到人们的关注. 如何为不同的学习者提供良好的个性化教学服务是虚拟学习社区研究的重要问题,而学习者的学习特征的提取与分析是个性化教学的基础. 以虚拟学习社区为背景,从学习者的学习过程和学习特征入手,运用模糊 c 均值聚类算法(FCM)挖掘并分析学习过程中学习者的学习特征,进而根据学习者的知识水平、自主学习和协作学习积极性等学习特征进行精确分组划分,以达有针对性的教学指导,实现个性化教学,提高学习者学习效率.

关键词:虚拟学习社区;模糊 c 均值聚类算法;学习特征;个性化教学
中图分类号:TP 391 文献标志码:A DOI:10.16357/j.cnki.issn1000-5862.2016.06.19

0 引言

当今社会是知识社会,是一个“知识爆炸”的社会,知识技术的快速更新迫使人们要不断地学习新的知识、新的技术,人们逐渐形成了终身学习的意识. 然而传统的教学方式已经满足不了人们追求终身学习的渴求,更无法解决发达地区与落后地区教学资源不均衡的问题. 伴随着计算信息技术的发展和 Internet 的普及,教学形式也发生了深刻的改变,网络教育迅猛发展,为新的教学形式提供了物质和技术支撑. 网络教育的快速发展,使得人们接受教育的方式也变得丰富多彩. 虚拟学习社区作为网络教育的一种新的发展方向,越来越受到人们的关注,人们对虚拟学习社区的网络支撑平台期望也越来越高.

虚拟学习社区不仅是一个学习者学习的网络平台,同时也是一个为各种学习特点的学习者提供教学服务的学习组织. 在虚拟学习社区中,存在着各种学习资源,学习者自身情况和学习特点,如知识水平、自主学习和协作学习积极性等都不尽相同,如何为不同的学习者提供良好的个性化教学服务是虚拟学习社区研究的重要问题. 而个性化教学可以提高虚拟学习社区学习资源的利用,个性化教学是指教师能够针对学习者的学习特征采取恰当的教学方法、策略指导学习者学习而并让学习者全面发展,而学习者的学习特征的提取与分析对个性化教学具有重要意义[1]. 同时在虚拟学习社区中的学习者特征的分析研究也有助于提高学习者的学习效果和社区的建设水平,因此对学习者特征进行研究显得十分必要[2-3]. 在虚拟学习社区中,个性化教学的内涵更丰富多彩,有协作学习、合作学习等,但始终离不开学习特征的分析. 在协作学习方面,王剑等[4]提出了个性化 e-Learning 协作学习推荐系统,将推荐系统和协作学习相结合,针对学习者的个性化特征,考虑学习者不同的学习能力,以更好地配合学习者之间的协作学习. 尹晨[5]通过分析学习者的学习特征,进行了协作学习的分组研究. 在学习者兴趣分组方面,程艳等[6]通过建立本体兴趣特征向量空间模型提出了虚拟学习社区的自组织算法. 李昕等[7]指出教学策略的制定应以学习者的学习特征为基础. 薛寒等[8]则根据每位学生的学习特征不同进行分班,制定个性化教学策略进行授课的施教,有助于学生的特长发展. 因为不同的学习者的兴趣爱好、知识水平、自主学习和协作学习积极性都不同,合适的分组方式将有利于解决学习者的个性化学习需求,可以较大地吸引学习者的学习兴趣,提高学习效率[9]. 本文从学习者的学习过程和学习特征入手,运用模糊 c 均值聚类算法(FCM),分析学习过程中的学习者的学习特征,并根据学习者的知识水平、自主学习和协作学习积极性等学习特征进行精确分组划分. 以达有针对性的教学指导,实现个性化教学,提高学习者学习效率的目的.

收稿日期:2016-10-02
基金项目:国家自然科学基金(61262080),江西省科技支撑计划重点项目(20151BBE50121)和江西省教育厅科技重点课题(GJJ15029)资助项目.
作者简介:程 艳(1976-),女,江西婺源人,教授,博士,主要从事虚拟社区和数据挖掘等方面的研究.

1　学习特征挖掘方法

1.1　数据预处理

通过虚拟学习社区平台记录学习者的学习动态,获取学习相关数据,再由数据预处理模块经过数据的清洗、数据集成、数据转换、数据规约等操作把杂乱、不规则、没有规律的数据转换成有用的数据.

1) 数据清洗: 数据清理的主要任务是补充缺失值,光滑噪音数据,识别或删除离群点以解决数据的不一致性. 学习者的学习特征是需要分析的数据对象,主要通过挖掘虚拟学习社区中的学习记录等数据来获取. 虚拟学习社区平台存储了大量学习者的学习信息,而学习信息又是一个多维的数据,在处理这些数据的过程中,由于机器或人为等多种原因,有些数据可能会偏离真实值,且会对分析的结果影响比较大,一般把这些数据当成离群点,进行清除处理.

2) 数据集成: 数据集成的目的是把来自不同地方的多个数据源的数据集中到一起. 由于数据来源不同,同一个数据或同一个数据的属性在不同的地方表达方式不一样. 为了避免数据的不一致性和冗余,对数据进行集成处理.

3) 数据规约: 即数据的简化处理. 比如学习者的平时成绩和期末成绩都是学习者知识水平方面的学习特征的一个反映,对其数据规约处理,用 1 维的知识水平综合表示平时成绩和期末成绩.

1.2　学习特征分析器

因为无论传统教育还是网络教育,知识水平、自主学习都是衡量学习者学习特征的典型指标; 另外,根据网络教育的特点,学习者的协作学习积极性将作为另外一个重要学习特征,故应综合考虑. 这里选取学习者的知识水平、自主学习、协作学习积极性 3 个方面来分析,并结合虚拟学习社区的实际教学情况,分别对知识水平、自主学习和协作学习积极性分析其影响因子及其权重.

定义 1　设有 n 个学习者,每个学习者的学习特征的维数为 m,则学习者学习特征向量为 $\boldsymbol{S}_i = (S_{i1}, S_{i2}, \cdots, S_{im})$. S_i 表示第 i 个学习者的学习特征向量,$\boldsymbol{S}_{im}$ 表示第 i 个学习者第 m 个学习特征值 $1 \leqslant i \leqslant n$.

本文对权重评价不作重点研究,其值的确定见文献 [10]. 其中知识水平由学习者的平时测验成绩和期末测验成绩衡量,并分别赋予相应的权重 $Q_1 = 0.3$,$Q_2 = 0.7$; 自主学习由作业资料模块练习次数、学习课程模块学习次数和测验模块测验次数进行评定,其权重分别为 $Q_3 = 0.3$,$Q_4 = 0.4$,$Q_5 = 0.3$; 协作学习积极性由资源共享模块学习次数和资料上传次数评定,其权重分别为 $Q_6 = 0.3$,$Q_7 = 0.7$. 在本文中,学习特征向量 $\boldsymbol{S}_i = (S_{i1}, S_{i2}, S_{i3})$,$S_{i1}$ 表示知识水平,S_{i2} 表示自主学习,S_{i3} 表示协作学习积极性,其具体学习特征的描述如表 1 所示.

表 1　学习者学习特征影响分析

学习特征	平时测验成绩(P_1)	期末测验成绩(P_2)	作业资料模块练习次数(T_1)	学习课程模块学习次数(T_2)	测验模块测验次数(T_3)	资源共享模块学习次数(X_1)	资料上传次数(X_2)
知识水平(S_{i1})	权重 Q_1	权重 Q_2					
自主学习(S_{i2})			权 重 Q_3	权重 Q_4	权重 Q_5		
协作学习积极性(S_{i3})						权重 Q_6	权重 Q_7

依托基于 Moodle 平台的江西师范大学计算机课程自主学习社区,本文对社区内学习者的学习行为记录进行采集、处理并挖掘分析. 该学习社中已开设了《大学计算机基础》等多门学习课程. 以 2014 级思想政治教育班的学习者为研究对象,对该班《大学计算机基础》的学习行为记录进行学习特征分析.

该班一共有 60 名学习者参与《大学计算机基础》课程的网络学习,根据上述数据处理方法,提取所需记录. 由于一些学生缺考等原因,共有 5 条无效记录,55 条有效记录. 根据上文中设定的各个影响因子的权重,计算出 55 名学习者相应的学习特征的值如表 2 所示.

表 2　部分学习者的学习特征

编号	知识水平得分	协作学习积极性 / 次	自主学习 / 次
1	83.91	1.7	42.1
2	40.00	2.2	12.8
3	83.33	2.9	7.7
4	95.65	1.0	23.1
5	88.51	9.9	67.9
6	90.90	1.9	8.2
7	53.89	1.4	31.1
8	44.44	7.2	14.7
9	59.88	4.0	9.1
10	66.29	0.7	2.8
…	…	…	…

2 基于模糊 c 均值的聚类分析

模糊 c-均值算法(FCM)是由 J. Bezdek [11] 提出的一种常用的模糊分类方法,FCM 算法可以将样本数据根据某种相似性划分为 c 个不同的类. 本文针对学习者的学习特征运用模糊 c 均值算法进行聚类分析,把学习特征相近的学习者归为一组,学习特征相异的划为另外一组,并分别求出每组学习者的聚类中心的特征向量. 通过对学习者学习特征的分析,由领域专家制定相应的教学策略,进行有针对性的教学,从而实现社区中学习者整体提高.

2.1 学习者聚类分析的问题描述

设有 n 个学习者,每个学习者的学习特征的维数为 m 维,学习者的集合为 $\boldsymbol{S} = \{S_1, S_2, \cdots, S_n\}$,学习者学习特征向量为 $\boldsymbol{S}_i = \{S_{i1}, S_{i2}, \cdots, S_{im}\}$. $\boldsymbol{S}_i$ 表示第 i 个学习者的学习特征向量,S_{im} 表示第 i 个学习者第 m 个学习特征值 $1 \leqslant i \leqslant n$.

学习者按学习特征分类后,每类学习者都有个聚类中心,分别形成一个聚类中心向量.

定义 2 学习者的学习特征经过聚类分析,把相似学习特征的学习者聚为一类,用每类的学习特征聚类中心向量 $\boldsymbol{Z}_k = (Z_{k1}, Z_{k2}, \cdots, Z_{km})$,$\boldsymbol{Z}_k$ 代表每类学习者的总体学习特征. Z_{km} 表示生成的第 k 类学习者的第 m 个学习特征值,k 为聚类类别数 c.

对学习者进行聚类分析也就是把学习者划分为 c 个类别. 学习者的每个分类结果对应着 $c \times n$ 阶学习特征隶属矩阵 $\boldsymbol{U} = \{u_{ij}\}$,$u_{ij}$ 为学习者与聚类子集 S_j 的隶属度. $\boldsymbol{V} = (v_1, v_2, \cdots, v_c)$ 为所有学习者 S_i 的聚类中心集合.

2.2 学习者聚类分析的目标函数

运用模糊 c 均值 [12] 根据学习特征对学习者进行聚类分析,算法的聚类的目标函数为

$$\min(J_m) = (\boldsymbol{U}, \boldsymbol{V}) = \sum_{i=1}^{c}\sum_{j=1}^{n}(u_{ij})^m d_{ij}^2, \quad (1)$$

要使得 J_m 达到最小,即每个学习者样本到聚类中心的距离平方和最短. 其中,$d_{ij} = \| \boldsymbol{x}_j - \boldsymbol{v}_i \|$ 表示第 j 个学习者与第 i 个聚类中心 $\boldsymbol{v}_i$ 之间的距离,其中 $\boldsymbol{x}_j$,$\boldsymbol{v}_i$ 分别表示具有 m 维学习特征向量的第 j 个学习者和第 i 个聚类中心点. m 为加权指数,通常取值为 2 [13],控制着隶属度的分配和聚类的模糊程度,取值越大模糊程度越高. u_{ij} 表示为学习者 S_j 对聚类中心 v_i 的隶属程度. 也就是说每个学习者都有可能隶属于每个类别,对每个类别都有一个隶属度. 同时 u_{ij} 必须满足以下条件:

(i) $u_{ij} \in [0,1]$,每个学习者对每个类别的隶属度在 0 ~ 1 之间;

(ii) $\forall i, \sum_{i=1}^{c} u_{ij} = 1$,每个学习者对所有的 c 个类别的隶属度之和为 1;

(iii) $\forall j, 0 < \sum_{j=1}^{n} u_{ij} < n$,对于某个类别,总存在至少一个学习者会隶属于它.

2.3 学习者隶属度及其聚类中心计算

学习者对每个类别的隶属度计算公式为

$$U_{ij} = \left(\sum_{k=1}^{c}(d_{ij}/d_{kj})^{2/(m-1)}\right)^{-1}, \quad (2)$$

其中 d_{ij} 表示第 j 个学习者到第 i 个聚类中心的距离,$\sum_{k=1}^{c} d_{kj}$ 表示第 j 个学习者到 c 个聚类中心点的距离之和. m 一般取值为 2.

学习者的聚类中心计算公式为

$$V_i = \sum_{j=1}^{n}(u_{ij})^m x_j / \sum_{j=1}^{n}(u_{ij})^m. \quad (3)$$

2.4 模糊 c 均值聚类分析步骤

模糊 c 聚类的算法步骤如下:(i) 模糊 c 均值聚类算法进行初始化设置. 确定学习者分类的个数 c,n 为学习者的个数,随机给定初始聚类中心 $V(0)$,设置迭代过程终止阈值 ε,设置初始迭代次数值 $p = 0$,(ii) 按(2)式计算隶属度 u 的值,得到 $\boldsymbol{U}^{(p)}$;(iii) 按(3)式计算新的聚类中心矩阵 $\boldsymbol{V}^{(p+1)}$;(iv) 利用(1)式计算 J_m,若 $|J_m^{(p)} - J_m^{(p-1)}| > \varepsilon$,则令 $p = p + 1$ 且转向(ii),否则停止计算出模糊隶属度矩阵 $\boldsymbol{U}$ 和聚类中心矩阵 $\boldsymbol{V}$.

3 实验过程

3.1 模糊 c 均值聚类算法的参数初始设置

1) 加权指数 m 的选取:加权参数 m 对算法的性能和效果影响较大. m 的选取既要保证类内加权误差平方和要小,又要保证类间距要大. 在本次实验中 m 取值为 2.

2) 最大迭代次数 p 的设置:迭代次数设为 150.

3) 迭代终止条件:设为 1×10^{-6},也即类内加权误差平方和的最小变化量小于 1×10^{-6},则停止迭代.

4) 聚类分组 c:聚类算法是种无监督的学习方法 [15]. 模糊 c 均值聚类算法中类别数 c 的选择应据具体情况具体分析. 本实验通过对比分析聚类类别数 c 与目标函数 J_m(学习者学习特征类内加权误差

平方和) 的关系,从而选择合适的学习者分组的类别数. 学习者类别 c 与 J_m 的曲线图见图 1. 由实验结果可知,随着 c 的增加,学习特征间的类内加权误差平方和 J_m 逐渐变小. 当 $c = 6$ 时,是个拐点,其后的变化较为平缓,故此处类别数取 3.

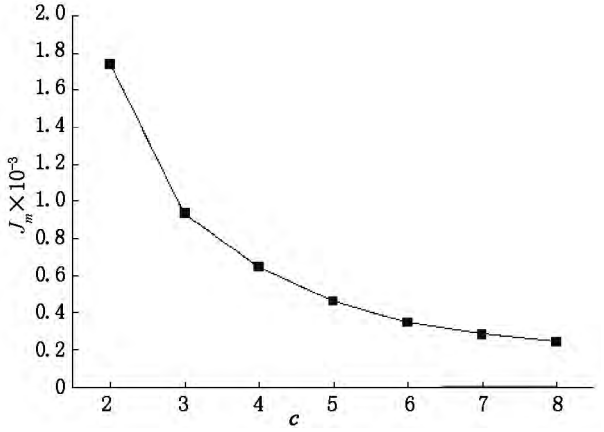

图 1　J_m 与聚类类别 c 变化关系图

3.2　模糊 c 均值聚类分组实现过程

1) 初始化学习特征聚类中心. 由上述分析可知,根据学习特征,学习者分组的合适组值为 3,本实验的初始聚类中心由系统中的函数随机选出 3 个学习特征聚类中心点.

2) 计算实际迭代次数. 本实验的最大迭代次数设置为 150 次,但当学习特征的最近的加权误差平方和的变化量小于设置的终止条件 0.000 01 时,则跳出迭代. 此次迭代次数为 51 次,满足终止条件.

3) 迭代 51 次后,确定学习者的学习特征最终聚类中心和最终隶属度划分,学习特征最终聚类中心和最终隶属度.

4) 根据隶属度把学习者划分成 3 个类. 每个学习者相对于 3 个聚类中心分别有个隶属度,学习者被划分到 3 个隶属度值最大的类别中.

算法终止后,模糊 c 均值聚类算法把学习者按学习特征划分成 3 类,如表 3 所示. 得到 3 组具有相似学习特征的学习者,其 3 组学习者的聚类中心学习特征向量.

表 3　聚类分组结果

聚类中心特征向量	自主学习/次	协作学习积极性/次	知识水平得分
中心特征向量 Z_1	72.860 4	3.841 1	81.148 4
中心特征向量 Z_2	10.743 4	2.092 2	19.463 7
中心特征向量 Z_3	14.302 1	3.976 4	51.205 8

由计算结果可知,根据学习者的知识水平、自主学习和协作学习积极性,通过模糊 c 均值聚类算法将该班 55 名同学进行分成了 3 组,从而将学习社区中一个相对较大的群体分成了具有相似特征的小群体,这样就可以进行有针对性的教学指导. 本文通过学习特征向量进行分组以后,由领域专家针对各个小组制定详细的教学策略和学习任务,并取得了较好的结果.

4　总结

本文以虚拟学习社区为背景,根据虚拟学习社区内的教学数据,提取学习者的学习特征,运用模糊 C 均值算法计算学习特征的聚类中心,把距离聚类中心距离近的学习特征分到同一类,从而把学习者也按照学习特征相近的分为一类. 最后根据学习特征由专家制定教学策略,进行个性化教学,为个性化教学提供服务.

5　参考文献

[1] 江吉林. 新媒体时代大学生网络学习特征分析 [J]. 软件导刊: 教育技术,2015(2):54-56.

[2] 刘砚. 网络学习者学习特征、影响因素及对策研究 [J]. 天津职业院校联合学报,2013(7):47-50.

[3] 张杰. 虚拟学习社区中学习者特征的分析研究 [J]. 电化教育研究,2008(12):67-71.

[4] 王剑,陈涛. 个性化 e-Learning 协作学习推荐系统研究 [J]. 中国远程教育,2016(7):1-9.

[5] 尹晨. e-Learning 协作学习中分组策略研究 [J]. 计算机技术与发展,2012(12):55-58.

[6] 程艳,许维胜,杨继君,等. 基于本体兴趣特征向量空间模型的社区自组织法 [J]. 系统工程,2009(5):96-103.

[7] 李昕,荆永君,王鹏. 智能授导系统中的教学策略研究 [J]. 中国电化教育,2012(10):126-130.

[8] 赵喜庆,王发成. 个性化教学促进一般生源特长发展研究 [J]. 现代中小学教育,2015(11):11-13.

[9] Viktor Freiman, Nicole Lirette-Pitre. Building a virtual learning community of problem solvers: example of CASMI community [J]. ZDM,2009,41(41):245-256.

[10] Bezdek J. Pattern recognition with Fuzzy Objective Function Algorithms [M]. New york: Plenum Press,1981.

[11] 孙晓霞,刘晓霞,谢倩茹. 模糊 c-均值(FCM) 聚类算法的实现 [J]. 计算机应用与软件,2008(3):48-50.

[12] 高新波,李洁,谢维信. 模糊 c 均值聚类算法中参数 m 的优选 [J]. 模式识别与人工智能,2000(1):7-11.

[13] 赵卫中,马慧芳,李志清,等. 一种结合主动学习的半监督文档聚类算法 [J]. 软件学报,2012(6):1486-1499.

(下转第 647 页)

covery,1998,2(2) : 121-167.
[11] Corinna Cortes,Vladimir Vapnik. Support-vector networks [J] . Machine Learning,1995,20(3) : 273-297.
[12] 邵坤艳. 基于视频图像的火灾检测方法研究 [D] . 重庆: 重庆大学,2015.
[13] 王林林. 基于机器视觉与图像处理技术的微钻刃面质量检测 [J] . 科技视界,2016(13) : 12-16.
[14] 马彩云. 基于图像处理技术的心率检测软件设计与实现 [J] . 山东工业技术,2016(11) : 88-92.
[15] 司红伟,全蕾,张杰. 基于背景估计的运动检测算法 [J] . 计算机工程与设计,2011,32(1) 262-273.
[16] 饶裕林. 基于视频的森林火灾识别方法研究 [J] . 电子世界,2016(10) : 79-83.

Based on Forest Fire Smoke Moving Object Detection Video Stream

BAI Shuhua,KUANG Mingxing
(Nanchang Institute of Technology,Nanchang Jiangxi 330044,China)

Abstract: Moving object detection is the precondition for video image classification and recognition. Smoke is a distinctive feature of the early forest fire,forest fire smoke through the characteristics of the image analysis,several common moving target detection methods,analyzes the implementation process,comparing their advantages and disadvantages,and to seek the best forest fire smoke video moving target detection methods. The experiments also showed that the improved method with color background estimation criterion method not only has a better ability to capture smoke and anti-jamming capability,which greatly reduces the subsequent image recognition pressure.

Key words: moving object detection; inter-frame difference method; background estimation; analyzing color

(责任编辑: 冉小晓)

(上接第 643 页)

[14] 程艳,苗永春. 高维数据流的聚类离群点检测算法研究 [J] . 江西师范大学学报: 自然科学版,2014,38(5) : 449-453.

Learning Behavior Feature Mining and Grouping Method for Virtual Learning Community

CHENG Yan[1,2],XIE Jianhua[1],TAN Pingfei[1],YANG Zhiming[1]
(1. College of Computer and Information Engineering,Jiangxi Normal University,Nanchang Jiangxi 330022,China;
2. Computer Science and Technology Post Doctoral Mobile Station,Tongji University,Shanghai 201804,China)

Abstract: Virtual learning community,as a new mode of network education,has attracted more and more attention. How to provide a good teaching service for different learners is an important issue in the virtual learning community,and the extraction and analysis of the characteristics of the learners are the basis of the individualized teaching. It starts from the learner´s learning process and learning characteristics,using the fuzzy *c*-means clustering algorithm (FCM) ,mining and analyzing the learning features of students in the learning process in the background of the virtual learning community. Students are accurate grouped according to the learner´s knowledge level,autonomous learning and cooperative learning initiative and other learning characteristics in order to achieve the goal of personalized teaching,to improve the learning efficiency of the learners.

Key words: virtual learning community; fuzzy *c*-means clustering algorithm; learning characteristics; personalized teaching

(责任编辑: 冉小晓)