

基于多分类器的迁移 Bagging 习题推荐

吴云峰, 冯 筠*, 孙 霞, 李 展, 冯宏伟, 贺小伟

(西北大学 信息科学与技术学院, 西安 710069)

(*通信作者电子邮箱 fengjun@nwnu.edu.cn)

摘 要:针对协同过滤(CF)推荐方法用户的历史信息不足等问题,提出基于多分类器的迁移 Bagging 习题推荐算法。主要思路是把推荐问题投入迁移学习框架,将待推荐习题的用户作为目标域,从中搜索相似历史信息的用户作为辅助域,帮助训练目标域以得到更准确的分类结果。实验结果表明,所提方法在习题推荐库及公开数据集上,比协同过滤算法性能提高了 10%~20%;比单分类器 Bagging 迁移算法性能提升了 5%~10%。该方法在一定程度上解决了习题推荐系统中存在的冷启动和数据稀疏问题,也可推广到商品推荐等电子商务平台。

关键词:迁移学习; Bagging; 协同过滤; 推荐系统; 计算机辅助教学

中图分类号:TP391.7 **文献标志码:**A

Online transfer-Bagging question recommendation based on hybrid classifiers

WU Yunfeng, FENG Jun*, SUN Xia, LI Zhan, FENG Hongwei, HE Xiaowei

(School of Information Science and Technology, Northwest University, Xi'an Shaanxi 710069, China)

Abstract: Traditional Collaborative Filter (CF) often suffers from the shortage of historic information. A transfer-Bagging algorithm based on hybrid classifiers was proposed for question recommendation. The main idea was that the recommendation and prediction problem were cast into the framework of transfer learning, then the users' demand for recommend questions were treated as target domain, while similar users who had applicable historic information were employed as auxiliary domain to help training target classifiers. The experimental results on both question recommendation platform and popular open datasets show that the accuracy of the proposed algorithm is 10%~20% higher than CF, and 5%~10% higher than single Bagging algorithm. The method solves cold start-up and sparse data problem in question recommendation field, and can be generalized into production recommendation on E-commerce platform.

Key words: transfer learning; Bagging; Collaborative Filtering (CF); recommendation system; Computer Assisted Instruction (CAI)

0 引言

随着 Internet 的发展,网络辅助教学系统(Online Aided Teaching System, OATS)为传统教育提供了随时随地的帮助。OATS 通常包括在线项目实践、在线习题测试、在线问答等功能。例如,通过在线教学平台,Sturm 大学法学院的各地考生能随时参加该院组织的一场专业考试^[1]。哈尔滨工业大学的 C 语言教学系统能完成在线考试、自动评分和统计工作^[2]。但是,目前大部分的 OATS 提供的只是简单辅助教学或者统一教学资源管理的功能,即给每个学生提供相同的习题和资源。然而不同学生的个性、背景不尽相同,对不同习题和知识点的理解力及偏爱也不一样。为了达到个性化教学、提高学习效率的目的,对不同学生推荐有针对性的习题和个性化的教学资源成为 OATS 系统的最新目标。

近几年,商品推荐系统广泛地应用在电子商务平台上^[3-5]。当前的在线推荐系统中,用户群被分为新用户(完全没有信息的用户)、信息量少的用户和老用户(背景信息很充足的用户)三类。传统的推荐算法只能应用在信息量少的用户和老用户上,其中,协同过滤(Collaborative Filtering, CF)是目前在商业营销中应用最成功的推荐模型^[6]。据报道,亚马

逊(Amazon)30%的销售额来自于它的推荐系统^[7-8]。

为了得到当前用户感兴趣的物品,协同过滤算法首先根据已有的少量信息计算两个用户之间的相似性;然后找出和当前用户相似性最高的老用户,即邻居;最后,根据邻居老用户的历史信息向当前较新用户进行商品推荐^[9-10]。因此对无历史信息的新用户群来说,协同过滤算法通常难以实施,这就是协同过滤算法的冷启动问题。此外,协同过滤算法还存在数据稀疏问题。这是由于商品信息较多,用户只对较少商品进行了评分造成的。例如在一个含有 10000 个商品的推荐系统中,新用户和老用户如果分别对了 10 个和 100 个商品评分,占有率仅为 0.001 和 0.01,因此新用户和老用户都不同程度地存在着数据稀疏的问题。

随着机器学习的快速发展,迁移学习框架给以上问题的解决提供了新思路^[11]。迁移学习具有通过相似的历史环境完成新环境的学习任务^[12-14]的能力。为了解决在线 Web 系统的冷启动问题, Pan 等提出了基于框架坐标迁移的矩阵分解^[15]。Kamishima 等提出了面向协同标签的迁移 Bagging 算法(Tr-Bagging)^[16]。但是,在 Tr-Bagging 算法中,作者只使用了简单的标签等级等特征,不能完整地描述用户信息,更无法直接应用在习题推荐系统中。

收稿日期:2013-01-28;修回日期:2013-03-08。

基金项目:国家自然科学基金青年基金资助项目(61202184);西北大学教改项目(ZC12020, JX12028);校级创新基金资助项目(2011059)。

作者简介:吴云峰(1988-),男,江西赣州人,硕士研究生,主要研究方向:个性化推荐、模式识别; 冯筠(1972-),女,陕西西安人,教授,博士生导师,主要研究方向:模式识别、医学图像处理、个性化教学; 孙霞(1977-),女,陕西西安人,副教授,主要研究方向:文本挖掘、信息检索。

本文提出了面向 C 程序网络辅助教学的习题推荐系统。该系统对学生的习题推荐采用两种方案:对于有丰富历史信息的学生,采用协同过滤算法推荐相关习题;而对于没有历史信息的新学生,本项目提出了一种基于混合多分类器的迁移学习算法。为了更好、更完整地描述用户信息,本项目提出了新的习题相关的特征组。实验表明,本项目提出的特征集合比文献[16]的特征集合更有效,取得了更好的习题推荐结果。

1 C 程序网络辅助教学系统

随着教学时间的压缩, C 程序设计的教学任务越来越难按时完成。为了帮助学生随时随地的进行程序设计的练习和实践,本项目发布了 C 程序网络辅助教学系统(Online C Programming Aided Teaching System, OCAT)。系统用户有两类,教师和学生,整个系统由三个功能模块组成,即信息收集模块、信息分析模块和习题推荐模块,简单描述如下。

1) 首先,在信息搜集模块中,由教师建立学生基本信息库、习题库、学生答题信息等数据库。学生通过登录和填表完善个人的扩展信息。完成登记之后,学生可获取推荐习题,查询答题成绩等,所有做过的习题及历史成绩均自动记入成绩数据库。例如:

①学生基本信息库中的某个学生信息。

学号:用作该学生的唯一标识;用户名/密码(用作登录之用);邮箱(用作找回密码之用)。

②学生扩展信息库中某个学生信息。

专业:软件工程;

爱好:篮球,竞技游戏,古典音乐;

年龄:18;

已会计算机技能:Word、Excel 和简单网页制作技术。

③习题库中某个习题信息。

习题编号:

题干:有一个正整数数组,包含 N 个元素,要求编程求出其中的素数之和;

类型:编程题;

难度系数:0.3;

已做学生百分比:67%。

④习题库中另一个习题信息。

题干:C 语言中,函数的隐含存储类别是_____。

A. auto

B. static

C. extern

D. 无存储类别

类型:选择题;

难度系数:0.1;

已做学生百分比:82%。

⑤学生答题库中的某个答题信息。

答题学号:12010111058

习题编号:01000772

答题时间:12 min

得分:A

2) 其次,信息分析模块完成和分析各种数据,例如学生兴趣爱好分析统计、习题的难易程度分析统计等。

例如:习题:

Int m = 2; int const * p1; p2 = &m; * p1 = 3;

Int * const p2 = &m; * p2 = 3; 请找出本题错误语句。

类型:选择题

知识点:指针

答对人数比例:73%

3) 最后,习题推荐模块负责推荐某学生可能感兴趣及易错的习题集。对于有历史背景信息的学生,推荐系统直接采用协同过滤算法实现;而对于没有多少背景信息的学生,本项目设计了基于迁移学习的习题推荐子系统,并提出了基于混合分类器的迁移 Bagging 的网络习题推荐算法。本文主要论述习题推荐模块。

2 基于迁移学习的习题推荐子系统

在习题推荐子系统中,设 $S = \{S_1, S_2, \dots, S_n\}$ 为学生集合, $I = \{I_1, I_2, \dots, I_m\}$ 为习题集合,其中: n 为学生数目, m 为习题数目。学生-习题的分数对集合 D 定义为三元组 $[S_i, I_j, Score_{(S_i, I_j)}]$ ($i \in [1, n], j \in [1, m]$), 其中 $Score_{(S_i, I_j)}$ 表示为学生 S_i 在习题 I_j 上的得分。为了简化习题推荐系统的复杂度,降低数据的稀疏性,定义习题得分等级 $R_k \in [1, q]$, 本文中 q 取 3, 即有 A/B/C 三个等级, 量化为:

$$R_{(S_i, I_j)} = \begin{cases} 1, & 0 \leq Score_{(S_i, I_j)} < 60 \\ 2, & 60 \leq Score_{(S_i, I_j)} < 80 \\ 3, & 80 \leq Score_{(S_i, I_j)} \leq 100 \end{cases} \quad (1)$$

则 $[S_i, I_j, Score_{(S_i, I_j)}]$ 转换为 $[S_i, I_j, R_k]$ 。学生 S_i 做过的习题表示为 $I_j(S_i)$ ($j \in [1, M(S_i)]$), $M(S_i)$ 为学生 $I_j(S_i)$ 做过的习题数目。

习题推荐子系统的主要任务是针对每一个学生进行习题推荐和评分。本文提出基于迁移学习的习题推荐子系统, 主要思想是首先将习题推荐问题纳入到机器学习的模式分类框架中, 将推荐问题转化为目标学生做某习题的等级分类问题。对于有丰富历史信息的同学, 传统的协同过滤算法能够解决相关习题的评分预测。对于没有历史信息的新学生, 提出了基于多分类器的迁移学习算法(Hybrid Transfer Bagging, HTR-Bagging)。首先计算历史信息的可迁移性, 再根据迁移性的优劣进行信息的迁移。本文还提出了和习题相关的得分比率及零分、满分率两组特征集合。实验表明, 本文提出的特征集合取得了很好的习题推荐结果, 基本解决了习题推荐系统中训练样本太少、分类器精度不足的问题。

2.1 目标域和辅助域的选择算法

对于每次习题推荐的分类任务, 首先以某个学生已做的习题集合及该学生的历史信息(学生-习题分数对)作为训练数据, 得到初始分类器, 并用此分类器完成未做习题(测试样本)是否给该生推荐的分类任务。然而, 传统监督机器学习算法需要大量已知标签样本进行分类器的训练。因此, 对于历史信息量不丰富(即没有做过多少习题)学生, 无法得到性能优良的强分类器。在本系统中, 如前所述, 习题推荐还存在固有的数据稀疏问题。对于那些历史信息非常少或者没有以往信息(即新用户的冷启动问题)的学生来说, 分类器甚至无法生成。本文首次提出利用迁移学习的思想解决习题推荐系统的冷启动和数据稀疏问题。

设目前针对某学生 S_i (历史信息习题信息很少或者没有历史信息)待推荐的习题集合为分类目标域 $D_T(S_i)$, 由于目标域的训练数据过少, 为了获得高性能的目标域分类器, 本文算法使用背景信息比较丰富的学生所做过的习题集合作为辅助目标域的训练数据 $D_A(S_i)$, 称为辅助域。辅助域选取的具体步骤如下。

1) 计算多数学生偏向选择的习题集合(即热门习题集合),并计算学生获分较多的等级集合(即热门等级集合 H^R)。

2) 找出选择热门习题集合的学生集合(即热门用户集合)。

3) 计算热门用户所选择的习题数据集合,作为辅助训练数据 $D_A(S_i)$ 。

定义针对学生 S_i 的进行推荐任务的混合数据集为 $D(S_i)$,即 $D(S_i) = D_T(S_i) \cup D_A(S_i)$ 。

2.2 习题样本的特征提取

为了适应模式分类的框架,需要提取 $D_A(S_i)$ 即辅助域和 $D_T(S_i)$ 即目标域中习题样本的特征向量。本文提出了两组特征向量集合,分别为 F_a 和 F_b 。第一组特征为 $F_a = \{R_1, R_2, \dots, R_q\}$,即衡量每道习题得分 q 个等级的学生数目。另一组特征

$$F_b = \{ZeroRatio, FullRatio\}, ZeroRatio(I_j(S_i)) = \frac{N(I_j'(S_i))}{N(I_j(S_i))}.$$

$$(score(S, I_j'(S_i)) = 0), FullRatio(I_j(S_i)) = \frac{N(I_j'(S_i))}{N(I_j(S_i))}.$$

$(score(S, I_j'(S_i)) = 100)$,其中: $(score(S, I_j'(S_i)))$ 中的 S 为所有学生, $N(I_j)$ 为做了习题样本 I_j 的学生数目。从上述定义可以看出, F_a 反映了某个习题的热门程度, F_b 则反映了某个习题的难易程度。

3 基于多分类器迁移 Bagging 的习题推荐算法

Bagging 算法是机器学习中能够有效提高分类精度的方法之一^[17]。它首先训练出一系列分类性能一般的弱分类器(基分类器),然后通过选择集成获得分类效果好的强分类器。传统的 Bagging 算法包括两个阶段,即弱分类器生成阶段和分类器选择集成阶段。目前文献中对分类器的选择主要是通过投票策略获得强分类器^[18-19]。本文提出了迁移 Bagging 算法,即在集成阶段将辅助域的分类器通过迁移算法,集成到目标域的分类结果中,借以提高目标弱分类器的分类性能。该算法最大限度地利用了辅助域的先验知识,有效地弥补了分类目标域的训练数据不足问题,为冷启动和数据稀疏提供了一个合理的解决方案。

针对传统迁移学习算法采用单分类器、对数据多样性建模不足的问题^[20],本文进一步提出了多分类器迁移 Bagging 算法(Hybrid Transfer Bagging, HTR-Bagging)。具体思想是在弱分类器生成阶段,采用多种不同类型分类器进行组合和选择集成,有效地提高了习题推荐集成分类器的性能。整个算法主要分为两个阶段,即多分类器训练阶段和多分类器迁移集成阶段。

3.1 多分类器训练

多分类器训练利用 2.2 节中得到的目标域数据集 $D_T(S_i)$ 和辅助数据集 $D_A(S_i)$ 。整个训练分为两个步骤:首先训练目标域数据集获得一系列弱分类器,作为第二步分类器集成的基础数据;然后在混合数据集 $D(S_i)$ 中随机抽取样本以得到有差异性的数据以进行进一步的训练。

设 $C = \{C_1, C_2, \dots, C_i, \dots, C_L\}$ 为不同类型的弱分类器,例如:KNN、Naïve Bayes、决策树等,其中 L 为基分类器的类别个数。同时设 T 为抽样的次数,每次抽样的样本数目为 $N_j(j \in \{1, 2, \dots, T\})$ 。 $CM_{(i,j)}$ 为第 j 次抽样获得的具体分类器模型(包括分类器类型和具体的分类器参数),其中对目标数据集训练的分类器模型为 $CM_{(i,0)}$, $CM_{(i,1)}$ 至 $CM_{(i,T)}$ 为混

合数据集上抽样得到的分类器模型。多分类器的训练算法的基本流程描述如下。

输入 $D_T(S_i), D_A(S_i), T, C, L$ 。

输出 $\{CM_{(1,0)}, CM_{(2,0)}, \dots, CM_{(L,0)}\}, \{CM_{(1,1)}, CM_{(2,1)}, \dots, CM_{(L,1)}\}, \{CM_{(1,T)}, CM_{(2,T)}, \dots, CM_{(L,T)}\}$ 。

步骤 1 目标域数据集 $D_T(S_i)$ 分类器训练。

采用 $D_T(S_i)$ 训练获得 $C = \{C_1, C_2, \dots, C_i, \dots, C_L\}$ 中各个分类器的参数,即获得分类器模型 $\{CM_{(1,0)}, CM_{(2,0)}, \dots, CM_{(L,0)}\}$ 。

步骤 2 混合数据集 $D(S_i)$ 的抽样分类器训练。

For $j = 1:T$

D_j 为从 $D(S_i)$ 中可重复随机抽样集合,数量为 N_j ,得到 $\{CM_{(1,j)}, CM_{(2,j)}, \dots, CM_{(L,j)}\}$

End for

3.2 多分类器迁移集成

3.1 节得到了 $(T+1) * L$ 个弱分类器,分类性能均在 50% 左右,并不能很好地预测习题推荐结果。在这些分类器中,前 L 个分类器是目标域的训练结果,由于训练数据少,因此性能更差。后面 $T * L$ 个分类器均为混合数据集的训练结果,分类器性能稍好。本文利用迁移学习和投票选择的原理,首先计算辅助域分类器在目标域上的迁移性能,再进行目标域和辅助域的选择集成,得到最终的强分类器。在分类器的集成过程中,根据辅助域向目标域的可迁移性能,确定辅助域分类器向目标域的迁移方案。具体的迁移性能通过计算辅助域分类器在目标域的分类结果平均绝对误差(Mean Absolute Error, MAE)来确定:

$$MAE = \left(\sum_{i=1}^{N_T} |p_i - r_i| \right) / N_T \quad (2)$$

其中: N_T 是数据集 $D_T(S_i)$ 的样本数目; P 为辅助域分类器在目标域上的测试结果; r 是目标学生做习题得分的真正的等级,即金标准。辅助域分类器在目标数据集 $D_T(S_i)$ 上的 MAE 越小,说明该分类器的迁移性能越好。

具体来说,多分类器的迁移集成的过程及伪代码如下:

输入 $D_T(S_i), T, L, \{CM_{(1,0)}, CM_{(2,0)}, \dots, CM_{(L,0)}\}, \dots, \{CM_{(1,T)}, CM_{(2,T)}, \dots, CM_{(L,T)}\}$ 。

输出 目标域强分类器。

步骤 1 每次抽样得到的 L 个分类模型中,选择出本次抽样在 $D_T(S_i)$ 上性能最好的模型 CM_* ,因此共获得 $(T+1)$ 个较好分类器模型,其中: $CM_{(*,0)}$ 为目标域分类器, $CM_{(*,1)}$ 到 $CM_{(*,T)}$ 为辅助域分类器。

For $j = 0:T$

从 $\{CM_{(1,j)}, CM_{(2,j)}, \dots, CM_{(L,j)}\}$ 选择 MAE 最小的分类器,即 $CM_{(*,j)}$

End For

步骤 2 对 CM_* 根据迁移性能 MAE 进行降序排序。得到有序的分类器组

$\{CM'_{(*,0)}, CM'_{(*,1)}, \dots, CM'_{(*,T)}\}$, 设置 $CM_{(*,j)}$ 在后续选择集成步骤的投票权重 CW 如式(3)所示:

$$CW = j / \sum_{i=1}^{T+1} i \quad (3)$$

步骤 3 利用选择集成中的投票集成方法,获得目标域中性能最好的分类器,用来完成后续该学生的习题推荐任务。

值得说明的是,本文提出的迁移训练得到的 $(T+1)$ 个分类器模型中,目标域模型的数目只有 1 个,其余 T 个模型为辅助域训练获得的模型。对于没有完全历史信息量的学生用

户,只是用后面 T 个辅助域分类器也可以进行习题推荐,因此对网络推荐系统中数据冷启动问题是一个较好的解决方案。

4 实验结果和分析

为了验证算法的有效性,本文分别在公开数据集 MovieLen^[11]和本项目发布的 OCAT 系统平台中搜集的数据集上进行测试。

实验中比较了本文提出的多分类器迁移学习算法(HTR-Bagging)和协同过滤(CF)算法^[2]。设 $L1$ 为用于 HTR-Bagging 的目标域用户的辅助用户数目, $L2$ 为用于 CF 的目标域用户的邻居数目。在两个算法的实验结果比较中,实验考虑了两种情况: $L1 = L2$ 及 $L1 \neq L2$ 。为了更好地测试多分类器的性能,实验中还对比了单分类器迁移 Bagging 算法和传统 Bagging 算法。基分类器则根据性能选用了 KNN 和 Naïve Bayes 两个较常见的传统分类器。OCAT 系统和算法对比实验都是在酷睿 i3 500, 2 GB DDR3 1333, Windows XP 平台下进行的,其中 OCAT 系统采用 VS2010 为开发平台,SQL Server 2000 为后台数据库存储;算法对比实验的平台为 Matlab 2010。

4.1 MovieLen 实验和结果

MovieLen 是一组来自电影推荐系统的数据集,是国际流行的测试协同过滤、迁移学习等算法的公开数据集^[11]。数据包括 943 个用户,1682 部电影,5 个等级($q = 1, 2, 3, 4, 5$),原始样本维数为 3,提取等级特征后,特征维数为 5。训练数据集的样本数为 80000,测试数据集的样本数为 20000,表 1 显示了每个等级的样本数目。

表 1 MovieLen 数据集不同等级的样本数目

等级	训练样本数目	等级	训练样本数目
1	4952	4	27354
2	8993	5	17033
3	21668		

从表 1 中可以看出,等级 3,4,5 的数目比其他的数目要多,因此选择它们为 H^R 。实验中,选择辅助训练集中用户数目分别为 $L1 = 4, 6, 8, 10, 12$ 。因只需在待推荐学生测试,所以测试集的数目和 $L1$ 相关。表 2 列出了 $L1$ 不同情况下的测试数据集的样本数目。

表 2 MovieLen 测试数据集的样本数目

$L1$	测试集数目	$L1$	测试集数目
4	593	10	1401
6	861	12	1737
8	1128		

图 1 显示当 $L1 = L2$ 时,各个算法推荐性能的对比情况。

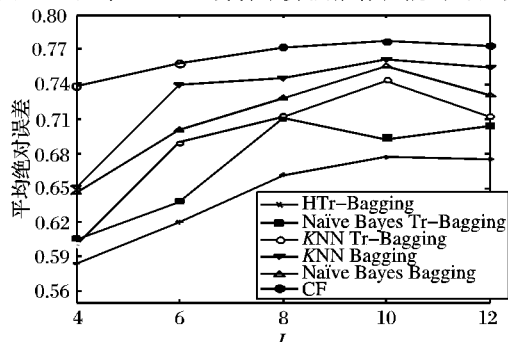


图 1 MovieLen 数据集上各算法的性能比较 ($L = L1 = L2$)

为了证明算法有效性,固定 $L1 = \{4, 12\}$, $L2$ 在 $\{4, 6, 8, 10, 20, 30\}$ 中变化。图 2 显示了固定 $L1 = 4$, $L2$ 变化的算法性能对比情况,图 3 显示了固定 $L1 = 12$, $L2$ 变化的算法性能对比情况。

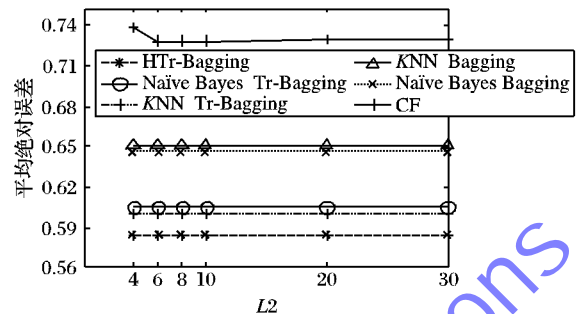


图 2 MovieLen 数据集上各算法的性能比较 ($L1 = 4, L2$ 变化)

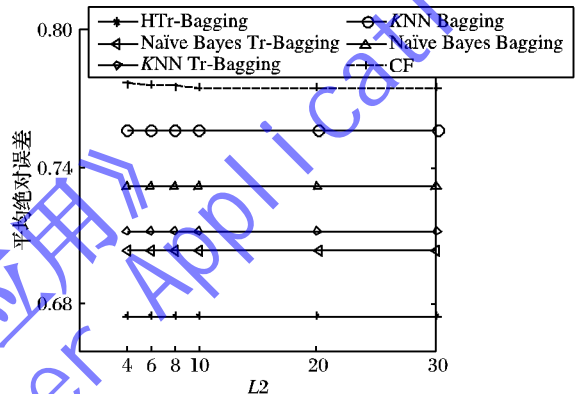


图 3 MovieLen 数据集上各算法的性能比较 ($L1 = 12, L2$ 变化)

从实验中结果,本文可以得出以下结论:基于多分类器迁移推荐算法(HTR-Bagging)比协同过滤算法(CF)得到更小的 MAE;迁移 Bagging 算法比 Bagging 算法得到更小的 MAE;本文提出 HTR-Bagging 算法比 CF, Bagging 算法和迁移 Bagging 算法的推荐性能更好。

4.2 OCAT 数据集实验和结果

OCAT 是本项目设计和实现的面向 C 语言教学的辅助教学系统。它记录学生在 C 语言习题上的历史成绩,并给学生进行习题推荐。该平台上采集的数据包括 547 个用户,1433 个习题,原始数据集样本个数为 47088,原始维数为 3,提取特征后维数为 5。数据样例如第 1 章所述。在接下来的实验中,应用 5 重交叉检验。表 3 为每个等级的训练样本的分布情况;表 4 列出了 $L1$ 不同情况下的测试数据集的样本数目。

表 3 OCAT 数据集不同等级的样本数目

等级	训练样本数目	等级	训练样本数目
1	14612	3	21534
2	1524		

其中:等级 1,3 被选中为 H^R 。

表 4 OCAT 数据集测试数据集数目

$L1$	测试集数目	$L1$	测试集数目
4	430	10	1044
6	631	12	1261
8	848		

在 2.2 节中提出了在 OCAT 平台上和习题相关的一组特征 $F_b = [FullRatio, ZeroRatio]$ 。为了验证此特征组的有效性,本文对比了包含和不包含 F_b 特征组的分类器性能。选择分

类器包括迁移 Bagging 和多分类器迁移 Bagging 算法。实验结果如图 4~6 所示。其中: HTr-Bagging(1) 只有特征集合 F_a , Tr-Bagging(2) 则使用 F_a 和 F_b 特征组合。

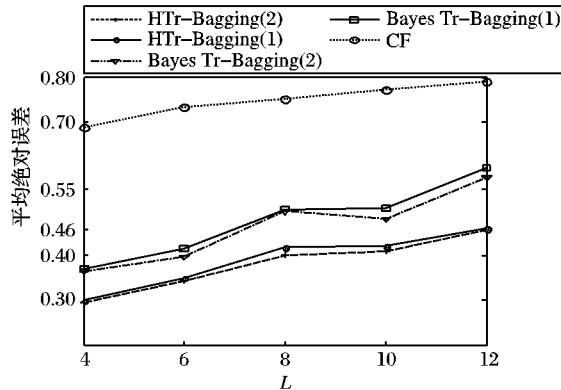


图4 OCAT数据集上各算法的性能比较($L = L1 = L2$)

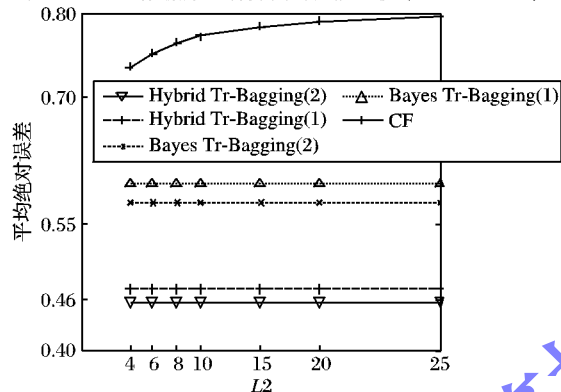


图5 OCAT数据集上各算法的性能比较($L1 = 4, L2$ 变化)

从实验结果中可以看出,在所有算法中,HTr-Bagging 算法得到最好的推荐效果;加入[FullRatio,ZeroRatio]两特征后,HTr-Bagging 比单组特征的推荐效果更好。

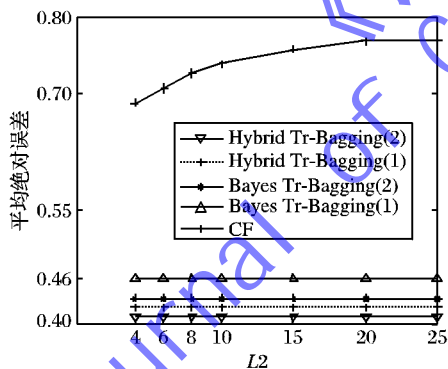


图6 OCAT数据集上各算法的性能比较($L1 = 12, L2$ 变化)

5 结语

针对习题推荐算法中的冷启动、数据稀疏等问题,本文提出多分类器 Bagging 迁移推荐算法。该方法创新地将习题推荐问题转化为迁移学习及模式分类框架中,进而使用辅助域丰富的历史信息帮助没有历史信息的目标用户完成推荐任务。从实验结果中可知,算法和传统协同过滤及目前流行的迁移学习算法相比,对新用户的习题推荐获得更好的性能。未来的工作包括更多特征集合的筛选及在更大规模的数据集上的测试。

参考文献:

[1] Online Exam System [EB/OL]. [2013-01-05]. <http://www.law.du.edu/forms/registrar/online-exams/>.

[2] 张华青,王红,滕兆明,等. 多维加权社会网络中的个性化推荐算法[J]. 计算机应用, 2011, 31(9): 2408-2411.

[3] RESNICK P, IACOVU N, SUCHA M, et al. GroupLens: an open architecture for collaborative filtering of netnews [C]// Proceedings of the 1994 ACM Conference on Computer Supported Cooperative Work. New York: ACM Press, 1994: 175-186.

[4] DESHPANDE M, KARYPIS G. Item-based top- N recommendation algorithms [J]. ACM Transactions on Information Systems, 2004, 22(1): 143-177.

[5] WANG J, SARWAR B. Utilizing related products for post-purchase recommendation in E-commerce [C]// Proceedings of the 5th ACM Conference on Recommender Systems. New York: ACM Press, 2011: 329-332.

[6] 刘建国,周涛,汪秉宏. 个性化推荐系统的研究进展[J]. 自然科学进展, 2009, 19(1): 1-15.

[7] 唐伟,周志华. 基于 bagging 的选择性聚类集成[J]. 软件学报, 2005, 16(4): 496-502.

[8] LINDEN G, SMITH B. Amazon.com recommendations: item-to-item collaborative filtering [J]. IEEE Internet Computing, 2003, 7(1): 76-80.

[9] HSU M-H. A personalized english learning recommender system for ESL students [J]. Expert Systems with Applications, 2008, 34(1): 683-688.

[10] DECEMMIS M, LOPS P, SEMERARO G. A content-collaborative recommender that exploits WordNet-based user profiles for neighborhood formation [J]. User Modeling and User-Adapted Interaction, 2007, 17(3): 217-255.

[11] CRAMMER K, KEARNS M, WORTMAN J. Learning from multiple sources [J]. Journal of Machine Learning Research, 2008, 9: 1757-1774.

[12] DAI W Y, YANG Q. Boosting for transfer learning [C]// Proceeding of the 24th International Conference on Machine Learning. New York: ACM Press, 2007: 193-200.

[13] About GroupLens [EB/OL]. [2013-01-05]. <http://www.grouplens.org/>.

[14] PAN S, YANG Q. A survey on transfer learning [J]. IEEE Transactions on Knowledge and Data Engineering, 2009, 22(11): 1345-1359.

[15] PAN W K, XIANG E W, LIU N N, et al. Transfer learning in collaborative filtering for sparsity reduction [C]// Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence. Atlanta: AAAI Press, 2010: 230-235.

[16] KAMISHIMA T, HAMASAKI M, AKAHO S. TrBagg: a simple transfer learning method and its application to personalization in collaborative tagging [C]// Proceedings of the 9th IEEE International Conference on Data Mining. Washington, DC: IEEE Computer Society, 2009: 219-228.

[17] BREIMAN L. Bagging predictors [J]. Machine Learning, 1996, 24(2): 123-140.

[18] LIU Q, GE Y, LI Z M, et al. Personalized travel package recommendation [C]// Proceedings of the 2011 IEEE International Conference on Data Mining. Washington, DC: IEEE Computer Society, 2011: 407-416.

[19] ISAACMAN S, IOANNIDIS S, CHAINTREAU A. Distributed rating prediction in user generated content streams [C]// Proceedings of the 5th ACM Conference on Recommender Systems. New York: ACM Press, 2011: 69-76.

[20] MIYAHARA K, PAZZANI M J. Collaborative filtering with the simple Bayesian classifier [C]// The Governance Body of Pacific Rim International Conferences on Artificial Intelligence, LNCS 1886. Berlin: Springer, 2000: 679-689.