基于谱聚类和 LFM 的选课推荐算法设计

刘旋

（信阳农林学院 信息工程学院，河南 信阳 464000）

摘　要：高校教务系统中学生数量和课程种类的飞速增长，使得传统推荐算法难以处理海量、高维的选课数据，为进一步提升大学生的选课效率，文章提出一种改进的 LFM 隐语义模型推荐算法，首先构造选课评分数据的相似矩阵，通过谱聚类进行初始分类，然后分类别构建 LFM 模型并计算合理的推荐算法。通过在某高校的选课数据集上的对比实验，证明了本文算法具有较高的预测精度和较低的空间复杂度。

关键词：推荐算法；隐语义模型；谱聚类算法

中图分类号：TP391　　**文献标识码：**A　　**文章编号：**2096-4706（2020）01-0014-03

A Recommended Courses Algorithm Based on Spectral Clustering and LFM

LIU Xuan

（College of Information Engineering，Xinyang Agriculture and Forestry University，Xinyang 464000，China）

Abstract：The rapid growth of the number of students and the types of courses in the educational administration system of colleges and universities，make the traditional recommendation algorithm is difficult to deal with mass and course of high-dimensional data，in order to further enhance students' course selection efficiency，this paper proposes a recommendation algorithm to improve the LFM argot meaning of model，the first data structure course score of similar matrix，the initial classification by spectral clustering and classification build LFM model and calculate the reasonable recommendations. Through the comparison experiment on the data set of course selection in a university，it is proved that the algorithm in this paper has higher prediction accuracy and lower space complexity.

Keywords：recommendation algorithm；LFM（latent factor model）；spectral clustering algorithm

0　引　言

人工智能时代的到来，使得高校教务系统逐渐成为大学生进行自选课的主要平台和途径，由于选修课程种类的增多，大多学生对课程缺乏了解且选择盲目。为了使学生更高效地选择与自身学习兴趣相似的课程，在传统教务选课系统中加入了推荐算法，过滤出符合学生要求的课程。目前，主流的推荐算法主要包括协同过滤、聚类推荐、奇异值分解等。例如文献 [1] 基于加权方式改进协同过滤算法，成功应用在学生课程推荐系统中，但忽视了数据的稀疏性；文献 [2] 通过计算选课数据间的相似性确定学生的近邻集合，采用概率矩阵分解的协同过滤算法进行推荐；文献 [3] 根据评论数据构建用户的偏好特征，并将相似偏好的用户聚类，然后使用 LFM[4] 模型（潜在因子模型，Latent Factor Model）进行推荐。

目前，诸多地方院校的选课推荐系统在处理海量选课信息时，存在推荐质量差、反馈时间长等问题，难以满足学习兴趣与教学效果的需求。LFM 算法作为一种隐语义模型方法，通过分析学生与课程的历史评分行为来寻找潜在联系，减少对其他因素的依赖，相比其他推荐算法具有逻辑简单、空间复杂度低且推荐结果更优等优点，非常适合处理大规模、高维的选课数据。但随着数据规模的扩大，LFM 算法的运算效率也大幅降低，且不同学生群体的选课兴趣也有较大差异，笼统地混为一谈进行计算会受到噪声数据的干扰，导致评分推荐结果不理想。

本文首先采用余弦距离作为相似度度量标准，计算学生选课评分的相似度矩阵，然后通过谱聚类方法对学生数据进行聚类，将学生群体进行划分，最后采用 LFM 算法分别构建不同类簇的评分数据矩阵，然后选取 Top-N 作为推荐结果。

1　相关理论

1.1　LFM 推荐算法原理

LFM 通过奇异值矩阵分解（SVD）处理评分矩阵，得到用户的潜在特征，以此评分缺失项目 [5]。对于高校学生的选课评分数据，LFM 认为课程评分值是学生对课程属性的喜好程度与每门课程在这些属性上的表现结果，并将评分矩阵 R 分解成为课程属性 P 和学生喜好 Q 两个矩阵的乘积，然后计算新的评分矩阵。

LFM 通过寻找隐藏因子将学生的喜好和课程属性联系起来。假设 R 是一个 $U\times I$（学生 × 课程）大小的评分矩阵，LFM 的核心思想是寻找两个低维矩阵 $P=U\times K$（课程的属性分类）和 $Q=I\times K$（学生的喜好分类），K 是隐因子的个数，通过 P 和 Q 的重新计算，得到新的预测评分矩阵 $R=P\times Q$，

收稿日期：2019-11-08

基金项目：河南省教育厅人文社会科学研究一般项目（2020-ZDJH-353）；信阳市哲学社会科学项目（2019JY048）

R 的表达式及损失函数分别如式（1）、式（2）所示：

$$R_{UI}=P_U Q_I=\sum_{K=1}^{K} P_{U,K} Q_{K,I} \tag{1}$$

$$C=\sum_{(U,I)\in K}\left(R_{UI}-\sum_{K=1}^{K} P_{UK} Q_{KI}\right)+\lambda\left(\|P_U\|^2+\|Q_I\|^2\right) \tag{2}$$

这里，$\|P_U\|^2+\|Q_I\|^2$ 是防止过度拟合的正则项系数，R_{UI} 为学生 U 对课程 I 的实际评分矩阵。通过计算损失函数 C 中两个参数的偏导数，根据式（3）的梯度下降法，通过不断迭代调整优化参数，找到最优的特征矩阵 P 和 Q，最后用 Q 和 P 两个矩阵的乘积，得到更新后的评分矩阵 R，以此预测的未评价的课程分数。

$$\begin{aligned}\frac{\partial C}{\partial P_{UK}}&=-2Q_{KI}\left(R_{UI}-\sum_{K=1}^{K} P_{UK} Q_{KI}\right)+2\lambda P_{UK}\\ \frac{\partial C}{\partial Q_{UK}}&=-2P_{KI}\left(R_{UI}-\sum_{K=1}^{K} P_{UK} Q_{KI}\right)+2\lambda Q_{UK}\end{aligned} \tag{3}$$

1.2　谱聚类

谱聚类是一种基于图论知识的降维聚类方法，具有计算量小、易于实现，善于处理高维数据的特点[6]。该方法首先构建基于相似度的无向权重图，然后按照切边规则将图分割为不同的子图，从而实现聚类，基于课程评分数据的谱聚类实现过程如下：

（1）输入 n 名学生对 m 门课程的评分样本集 $D=\{x_{ij}, i=1,\cdots,n; j=1,\cdots,m\}$，设定学生评分数据 $S'=\{x_1,\cdots,x_n\}$ 间的相似性度量标准，以此构造相似矩阵 S；

（2）根据相似矩阵 S 构建邻接矩阵 W 以及度矩阵 O；

（3）计算拉普拉斯矩阵 $L=O-W$，并计算出 L 的前 k 个最小特征值所对应的特征向量：$u_1, u_2, \cdots, u_k$；

（4）将上面的 k 个列向量组成 $n\times k$ 矩阵并标准化：$V=\{u_1, u_2, \cdots, u_k\}$；

（5）使用 K-means 算法进行聚类，得到类簇集合 $B=\{b_1, b_2, \cdots, b_h\}$（$h$ 为聚类个数）。

由于谱聚类算法是基于降维的方法，所以更适用于高维度数据的处理，而且无需考虑样本空间的形状，仅需计算数据集间的相似度矩阵就能实现聚类，相比传统聚类方法在处理高校课程评分的高维、稀疏数据时更加有效。

2　基于谱聚类的 LFM 推荐算法

针对学生的部分课程分数据集的高维性与稀疏性，本文首先计算数据间的相似度矩阵，然后采用谱聚类进行初始分类，使得具有相同选课兴趣的学生聚集到一起，最后根据不同兴趣的群体分别采用 LFM 模型进行评分预测，并从中选取 Top-N 门课程推荐给相应的学生。

2.1　谱聚类过程

根据数据维度高、数值稀疏等特点，本文采用余弦距离度量学生评分间的相似性。假设学生 i 和学生 j 对评分课程的相似度表示为：

$$sim(i,j)=\frac{\sum_{l\in T_{ij}} r_{li}\cdot r_{lj}}{\sqrt{\sum_{l\in T_j} r_{lj}\cdot\sum_{c\in T_i} r_{li}}} \tag{4}$$

其中：T_i 和 T_j 分别为学生 i 和 j 的评分课程集合；T_{ij} 为学生 i 和 j 所有的评分课程；r_{li} 与 r_{lj} 分别为学生 i 和 j 对课程 l 的评分。

基于相似度权重采用全连接法构造初始的相似性矩阵 S，以此作为 1.2 节谱聚类的初始输入，然后利用 K-means 聚类得到 n 个选课兴趣类别的学生群体。

2.2　LFM 推荐过程

LFM 模型通过输入课程评分矩阵，计算得到学生对课程隐藏兴趣分类的喜好程度。由于谱聚类分解了初始评分矩阵的规模，且每一类的数据相似度更高，提升了 LFM 的抗噪性，使得 LFM 的推荐结果更准确，具体步骤如下：

（1）随机初始化评分矩阵 P 和 M 的值；

（2）根据式（3）计算 P 和 M 的梯度 ∇P 和 ∇M，确定最快的下降方向；

（3）然后通过式（4）计算新的 P_{UK} 与 Q_{UK}，其中 α 表示学习速率，其值越大，则迭代下降得越快，λ 为正则化参数，用于防止过拟合。α 和 λ 均可通过实验统计得到：

$$\begin{aligned}P_{UK}&=(1-\lambda)\ P_{UK}+\alpha\left(R_{UI}-\sum_{K=1}^{K} P_{UK} Q_{KI}\right)Q_{KI}\\ Q_{UK}&=(1-\lambda)\ Q_{UK}+\alpha\left(R_{UI}-\sum_{K=1}^{K} P_{UK} Q_{KI}\right)P_{KI}\end{aligned} \tag{5}$$

（4）循环迭代步骤（2）（3），直到开始收敛或设置指定的迭代次数时停止；

（5）根据得到的最优 P 和 M 计算出最终的评分预测矩阵。

2.3　改进算法的详细步骤

输入数据：m 名学生对 n 门课程的评分表 R（存在大部分缺失值）。

输出数据：m 名学生的推荐课程列表。

Step1　根据评分表 R，采用式（4）计算 m 名学生之间的相似度矩阵 S。

Step2　根据 S 谱聚类得到 B 个学生类簇。

Step3　初始化 B 个矩阵集合 $\{P_1, P_2, \cdots, P_C\}$ 和 $\{Q_1, Q_2, \cdots, Q_C\}$，采用 LFM 计算得到新的评分预测矩阵 R。

Step4　基于 Top-N 算法从 R 中选择分值较大的 k 门课程推荐给相应学生。

3　实验比较

3.1　数据来源

本文的实验数据来源于某大学教务管理 Web 系统中学生选课数据集，共 1057 名计算机专业学生，15 门选修课程，15687 条在 2018 学年的选课评分信息，总共包含 600 条课程评分，每名学生至少对 6 门课程进行了评分。

3.2　实验评测标准

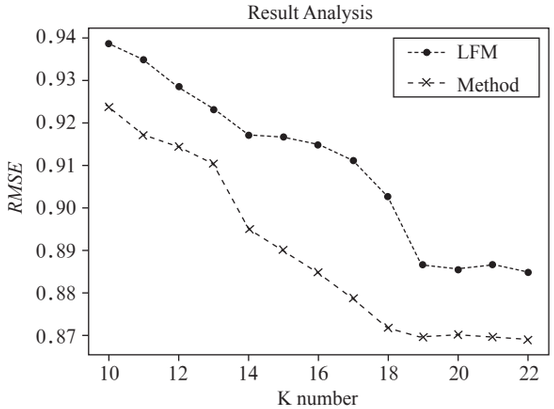
本文实验中使用均方根误差（*RMSE*）和时间作为主要评价标准，均方根误差项又可称为标准误差，它是用来反映一个数据的离散程度。

$$RMSE=\sqrt{\frac{\sum_{i=1}^{n}\sum_{j=1}^{m}(R_{ij}^{'}-R_{ij})^2}{N}} \tag{6}$$

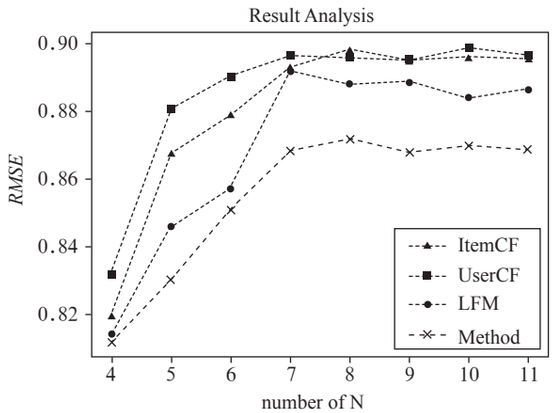
式中，R_{ij} 为推荐算法计算出的学生 i 对选修课 j 的预测评分，R'_{ij} 为教务系统中学生 i 对选修课 j 的实际评分，N 为数据集中的课程评分数。

3.3　实验分析

本文实验环境为：Python3.6、Win10 系统、内存 16G、处理器 i7-7400。针对该校的选课评分统计数据，将本文算法与 UserCF[7]、ItemCF[7] 和 LFM 算法进行试验对比。



（a）隐因子个数影响



（b）推荐项目数影响

图 1　不同参数下的实验结果

首先分析 LFM 矩阵分解过程中隐因子个数 K 对算法的影响，观察图 1（a）可知，当推荐项目数固定时，随着 K 的增大，LFM 和本文算法的 *RMSE* 均逐步降低，当 $K>15$ 时，*RMSE* 逐渐趋于稳定，且本文算法较 LFM 误差降低。其次测试 Top-N 推荐中，推荐课程数目对推荐结果的影响，对比算法在不同 N 值下的误差如图 1（b）所示，各算法随着推荐数目的增加，误差逐渐提高，当 $N>8$ 之后，*RMSE* 开始趋于稳定。在评分准确率上，本文算法在不同的推荐数目上，均取得最优，说明通过谱聚类，不仅分解了数据规模，而且优化了 LFM 的初始评分矩阵，在一定程度上降低了噪声干扰，从而使推荐结果更精确。

4　结　论

本文在 LFM 的基础上融合了谱聚类，改进算法在推荐过程中需要考虑到相似学生间选课的关联性，通过聚类优化 LFM 的初始评分矩阵，从而提高算法的推荐准确性。如何增量式地处理动态选课评分数据，是本文未来的研究重点。

参考文献：

[1] 沈苗，来天平，王素美，等 . 北京大学课程推荐引擎的设计和实现 [J]. 智能系统学报，2015，10（3）：369-375.

[2] 陈万志，张爽，王德建，等 . 基于近邻模型与概率矩阵分解的高校选课推荐算法 [J]. 辽宁工程技术大学学报（自然科学版），2017，36（9）：976-982.

[3] GANU G，KAKODKAR Y，MARIAN A. Improving the quality of predictions using textual information in online user reviews [J]. Information Systems，2013，38（1）：1-15.

[4] KOOPMAN S J，LUCAS A，MONTEIRO A A. The Multi-State Latent Factor Intensity Model for Credit Rating Transitions [J]. SSRN Electronic Journal，2008，142（1）：399-424.

[5] 陈晔，刘志强 . 基于 LFM 矩阵分解的推荐算法优化研究 [J]. 计算机工程与应用，2019，55（2）：116-120+167.

[6] DHILLON I S，GUAN Y，KULIS B. Kernel k-means：spectral clustering and normalized cuts [C]//KDD’04 Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining，Seattle，USA，August 22-25，2004. New York：ACM，2004：551-556.

[7] WANG J，VRIES A P D，REINDERS M J T. Unifying user-based and item-based collaborative filtering approaches by similarity fusion [C]//SIGIR 2006：Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval，Seattle，Washington，USA，August 6-11，2006. ACM，2006.

作者简介：刘旋（1991–），男，汉族，河南信阳人，讲师，硕士，研究方向：数据挖掘、模式分析。

（上接 13 页） 据爬虫程序设计 [J]. 信息系统工程，2016（9）：97-99.

[7] 陆树芬 . 基于 Python 对网络爬虫系统的设计与实现 [J]. 电脑编程技巧与维护，2019（2）：26-27+51.

[8] 熊畅 . 基于 Python 爬虫技术的网页数据抓取与分析研究 [J]. 数字技术与应用，2017（9）：35-36.

[9] 吴爽 . 基于 python 语言的 web 数据挖掘与分析研究 [J]. 电脑知识与技术，2018，14（27）：1-2.

作者简介：温娅娜（1999.03–），女，汉族，内蒙古包头人，本科，学士学位，研究方向：人工智能和软件开发应用。