基于 FCM 用户聚类的协同过滤推荐算法

赵学健[1],张雨豪[1],陈 昊[1],刘 旭[2],李朋起[3]
(1. 南京邮电大学 现代邮政学院,江苏 南京 210003;
2. 南京邮电大学 通信与信息工程学院,江苏 南京 210003;
3. 南京邮电大学 物联网学院,江苏 南京 210003)

摘 要:传统的协同过滤推荐算法存在数据稀疏性以及推荐准确率低等问题,针对该问题提出一种基于模糊 C 均值聚类的协同过滤推荐算法 GAFCM-CF(genetic algorithm based fuzzy c-means collaborative filtering)。首先,该算法结合用户评分和项目特征构建用户特征偏好矩阵,深入挖掘利用用户隐藏信息。其次,该算法通过模糊 C 均值聚类算法对用户进行聚类,并且为了防止模糊 C 均值聚类算法收敛于局部极小值,影响推荐质量,该算法基于遗传算法对模糊 C 均值聚类算法进行了改进,防止出现局部最优解。最后,该算法综合考虑了用户特征偏好矩阵以及用户项目评分矩阵计算用户相似度,实现推荐。实验结果表明,所提出的基于改进模糊 C 均值聚类的协同过滤推荐算法相比于传统的基于用户的协同过滤推荐算法及 PDSFCM 算法具有更好的推荐质量,提高了推荐的准确率。

关键词:推荐算法;协同过滤;模糊 C 均值聚类;遗传算法;评分矩阵
中图分类号:TP391 **文献标识码**:A **文章编号**:1673-629X(2021)08-0006-07
doi:10.3969/j.issn.1673-629X.2021.08.002

Collaborative Filtering Recommendation Algorithm Based on Fuzzy C-Means User Clustering

ZHAO Xue-jian[1],ZHANG Yu-hao[1],CHEN Hao[1],LIU Xu[2],LI Peng-qi[3]
(1. School of Modern Posts,Nanjing University of Posts and Telecommunications,Nanjing 210003,China;
2. School of Telecommunications & Information Engineering,Nanjing University of Posts and Telecommunications, Nanjing 210003,China;
3. School of Internet of Things,Nanjing University of Posts and Telecommunications,Nanjing 210003,China)

Abstract:Aiming at the problem of data sparsity and low accuracy of traditional collaborative filtering recommendation algorithms,a new genetic algorithm based fuzzy c-means collaborative filtering recommendation algorithm named GAFCM-CF is proposed. Firstly,the user feature preference matrix is constructed based on user rating matrix and item characteristics,and the hidden information of users is deeply mined. Secondly,the fuzzy c-means clustering algorithm is used to cluster the users. In order to prevent the fuzzy c-means clustering algorithm from converging to the local minimum and affecting the recommendation quality,the proposed algorithm improves the fuzzy c-means clustering algorithm based on genetic algorithm to prevent the local optimal solution. Finally,the user similarity is calculated by considering both the user characteristic preference matrix and user item rating matrix to realize better recommendation. The experiment shows that the proposed collaborative filtering recommendation algorithm based on improved FCM has better recommendation quality and improves the accuracy of recommendation compared with the traditional user-based collaborative filtering recommendation algorithms and PDSFCM algorithm.

Key words:recommendation algorithm;collaborative filtering;fuzzy c-means clustering;genetic algorithm;rating matrix

0 引 言

信息技术和互联网技术的迅猛发展,使得数据量呈指数性爆炸,人民逐渐从信息匮乏的时代走入了信息过载的时代[1]。无论是信息生产者还是销售者都遇到了很大的挑战,对于消费者而言,海量的数据筛选,获取有效信息越来越困难;生产者为了满足客户需求,生产有价值的信息,变得越来越困难。推荐算法是一种有效的信息处理工具,通过用户的历史行为信息,将

收稿日期:2020-06-08 **修回日期**:2020-10-10
基金项目:国家自然科学基金项目(61972208);中国博士后科学基金(2018M640509);南京邮电大学项目(NY217028)
作者简介:赵学健(1982-),男,博士,副教授,CCF 会员(88401M),研究方向为数据挖掘、无线网络关键技术。

用户和商品联系起来,解决信息过载的问题。目前,推荐算法已经成功应用到电子商务、在线音视频网站以及社交网络平台等各个领域。亚马逊的前首席科学家 Andreas Weigend 提及亚马逊有 20% ~30% 的销售来自于推荐系统[2]。

推荐算法是推荐过程的重要组成部分,为推荐系统的核心内容。目前有许多种推荐算法,常见的推荐算法有基于人口学的推荐算法、基于内容的推荐算法、基于关联规则推荐算法、协同过滤推荐算法、混合推荐算法。而协同过滤推荐算法是目前发展最为成熟、应用最为广泛的个性化推荐技术之一。协同过滤算法可以分为基于内存(memory - based)的和基于模型(model-based)的两类[3]。其中基于内存的协同过滤推荐算法又可以分为基于用户的协同过滤算法和基于项目的协同过滤算法。

1　研究现状

随着电子商务深入人心,用户和项目的数量急剧增加,这使得协同过滤推荐算法计算量巨大,时间复杂度和空间复杂度都极大。另一方面,单个用户所关注的项目通常都很少,这又导致用户的评分矩阵极其稀疏,使得推荐系统的精度大大降低。近年来,研究者开始借助聚类方法来解决协同过滤推荐过程中的数据稀疏性和推荐精度降低的问题。

文献[4]提出了一个新的基于 Web 的推荐系统,该系统基于用户在 Web 页面上浏览的顺序信息,采用模糊 C 均值聚类算法为目标用户确定相似用户,并评估每个网页的权重,来预测推荐用户的下一次访问网页,极大提高了现有推荐系统的精度。

文献[5]提出一种用于医学图像模糊聚类与直觉模糊推荐结合的混合推荐模型 - HIFCF(hybrid intuitionistic fuzzy collaborative filtering)。该模型比传统的模糊集合或单纯的推荐系统具有更好的预测精度。

文献[6]提出一种新的社交推荐模型,该模型首先将描述多个领域用户偏好的用户偏好矩阵形式化,然后利用偏距离策略模糊 C-均值聚类算法-PDSFCM(partial distance strategy fuzzy c-means)得到用户聚类分组,然后设计了一个基于聚类的社交正则化项,将聚类关系与传统的矩阵分解模型进行融合,用以进一步提高推荐算法的精度。

文献[7]提出一种新的基于聚类的协同过滤方法-CBCF(clustering-based collaborative filtering),该方法基于用户评分数据建立激励/惩罚用户模型,对用户进行聚类,在不需要更多先验信息的情况下,提高了推荐的准确性。

文献[8]将单领域基于聚类的矩阵分解方法扩展应用到多领域推荐,所提出的推荐方法可以更有效地利用来自辅助域的数据来获得更好的推荐效果,特别是对于冷启动用户。

文献[9]在 2010 年通过提出一种基于用户偏好模糊聚类的协同过滤推荐,用以解决推荐过程中的数据稀疏性和伸缩性。该方法将用户项评分矩阵转换为用户类矩阵,因此大大提高了矩阵中数据的密度。然后,使用模糊 C 均值算法将用户模糊地分为不同的组。采用模糊 C 均值聚类可以让每个用户属于不同的组,可以更为有效地捕获用户的各种偏好。

文献[10]在 2015 年提出了一种结合 FCM 和 Slope One 算法[11]的协同过滤推荐方法,该方法针对推荐算法的数据稀疏性问题,首先使用基于 FCM 聚类的 Slope One 算法来预测未评分的数据,然后通过基于用户的协同过滤推荐算法来实现推荐。

文献[12]为了提高推荐质量,将信任关系融合到推荐系统中,采用模糊 C 聚类算法,对信任关系进行聚类。利用信任类预测用户间的隐式信任,最后将信任关系与用户-项目关系线性融合进行推荐。实验表明该算法能够大幅度地改进推荐质量,提升算法的时间效率。

文献[13]为了克服评级数据的稀疏性问题,提出了一种新颖的稀疏性消除方法,该方法结合了评级和电影题材特征,应用模糊 C 均值聚类技术对电影进行聚类。该方案结合了评分和电影的题材来预测未评分数据,有效提升了推荐质量。

文献[14]提出了一种基于对用户真实性信息应用模糊 C 均值聚类的协作过滤模型。该文献提出一种新的度量用户相似度的方式,该公式结合了用户的使用组合系数对模糊真实性信息进行评级,在数据稀疏和冷启动条件下,推荐效果更佳。

文献[15]针对推荐算法的数据稀疏性和冷启动问题,将聚类算法和关联规则生成算法相结合,首先根据用户相似度对评分矩阵进行聚类,然后将聚类数据转换成布尔数据,并生成高效的关联规则,最后进行基于规则的推荐。实验表明,该方法不仅降低了推荐系统的稀疏度,而且提高了推荐系统的精度。

通过上述分析,可以看出当前借助聚类方法的协同过滤推荐通常只考虑了用户的显性特征进行聚类,没有考虑到项目的隐性特征;另一方面,当前采用模糊 C 均值聚类方法对用户进行聚类时,该算法容易收敛于局部极小值点,有时难以取得目标函数的全局最小值。因此,该文提出一种基于 FCM 用户聚类的协同过滤推荐算法 GAFCM-CF(genetic algorithm based fuzzy c-means collaborative filtering)。该算法首先结合用户

评分和项目特征构建用户特征偏好矩阵,然后采用模糊C均值聚类算法对用户进行聚类。此外,该算法为了防止模糊C均值聚类算法收敛于局部极小值,影响推荐质量,采用遗传算法对模糊C均值聚类算法进行了改进,以防止模糊C均值聚类算法出现局部最优解。实验结果表明,所提出的基于改进FCM的协同过滤推荐算法GAFCM-CF相比于传统的基于用户的协同过滤推荐算法具有更好的推荐质量。

2 算法理论基础

2.1 基于用户的协同过滤推荐算法

基于用户的协同过滤算法是推荐系统中比较古老的推荐算法,这个算法的诞生标志着推荐算法的诞生。该算法利用目标用户的历史行为信息,挖掘与目标用户具有高相似度的近邻用户集合,然后根据用户对此项目的评分来预测目标用户对该商品的相应的评分,之后再从预测的评分中选择靠前的Top-K个项目推荐给用户。

基于用户的协同过滤算法中,用户-项目评分矩阵 $\boldsymbol{R}^{m\times n}$ 是算法的基础,如表1所示。该矩阵中,每行对应一个用户,每列对应一个项目,每个矩阵元素 $r_{i,j}$ 表示用户 i 对项目 j 的评分,当用户没有对项目进行评分时,$r_{i,j}$ 为0或者NULL。

表1 用户项目评分表

	Item_1	Item_2	…	Item_n
User_1	r	r_{12}	…	r_{1n}
User_2	r_{21}	r_{22}	…	r_{2n}
…	…	…	…	…
User_m	r_{m1}	r_{m2}	…	r_{mn}

在基于用户的协同过滤推荐算法中,可以选择皮尔逊相关系数、余弦相似度等不同的相似度计算方法。皮尔逊相关系数计算方法如公式(1)所示:

$$\text{Sim}(u,v)=\frac{\sum_{i\in I_{u,v}}(r_{u,i}-\overline{r_u})(r_{v,i}-\overline{r_v})}{\sqrt{\sum_{i\in I_{u,v}}(r_{u,i}-\overline{r_u})^2}\sqrt{\sum_{i\in I_{u,v}}(r_{v,i}-\overline{r_v})^2}} \tag{1}$$

式中,$I_{u,v}$ 表示用户 u 和用户 v 共同评分的商品构成的集合;$r_{u,i}$ 是用户 u 对商品 的评分;$\overline{r_u}$ 表示用户 u 所有评分的平均值。

2.2 模糊C均值聚类算法

模糊C均值聚类算法(fuzzy c-means,FCM)是在硬C均值聚类算法模型基础上融合了模糊理论的精髓进一步推理得到的。硬C均值聚类算法要求每个用户只能明确属于某一个类之中,然而模糊C聚类可以提供更加灵活的聚类结果,它可以将每一个目标对象划分到多个类中。

假设数据集 $X=\{x_1,x_2,\cdots,x_n\}\subset R^{d\times n}$,其中 n 为数据集的个数,d 为数据集的维度。模糊C均值聚类算法将数据集划分成 k 个子集,则对应生成模糊划分矩阵 $\boldsymbol{U}$,$c_j(j=1,2,\cdots,k)$ 为每个聚类的中心,可记录为 C,$\mu_{i,j}$ 是第 i 个样本对应第 j 类的隶属度函数,则基于隶属度函数的聚类损失函数如公式(2)所示:

$$J_f=\sum_{j=1}^{n}\sum_{i=1}^{c}\mu_{i,j}^{m}\parallel x_i-c_j\parallel^2$$
$$\text{s.t.}\ \sum_{i=1}^{c}\mu_{i,j}=1,\forall j=1,2,\cdots,n \tag{2}$$

其中,m 是加权指数,也可以称为平滑系数,一般取值为2。

模糊C均值聚类算法首先计算各个用户和聚类中心之间的距离,然后计算出用户对各聚类中心的隶属度矩阵,通过比较用户在各个聚类中心隶属度的大小,将用户分配到隶属度最大的用户簇中,使得在同一个用户簇之中用户与用户的相似度最高,降低不同用户簇中用户之间的相似度。使得聚类函数最小的必要条件为 c_j 和 $\mu_{i,j}$ 分别满足公式(3)和公式(4):

$$\mu_{i,j}=\left[\sum_{k=1}^{c}\left(\frac{\parallel x_i-c_j\parallel}{\parallel x_i-c_k\parallel}\right)^{\frac{2}{m-1}}\right]^{-1}$$
$$1\leqslant i\leqslant n,1\leqslant j\leqslant c \tag{3}$$

$$c_j=\frac{\sum_{i=1}^{n}u_{i,j}^{m}x_i}{\sum_{i=1}^{n}u_{i,j}^{m}} \tag{4}$$

3 GAFCM-CF算法

该文提出的GAFCM-CF算法包括数据预处理,用户特征偏好矩阵构建,矩阵归一化处理,GAFCM聚类,用户相似度计算,目标项目评估及推荐六个步骤,如图1所示。算法的核心是用户特征偏好特征矩阵的构建和融合遗传算法对模糊C均值聚类算法进行改进,实现对用户的聚类分析,防止模糊C均值聚类算法出现局部最优解。

3.1 数据预处理

数据预处理主要负责从原始数据中提取用户特征和项目特征数据并进行数据清洗操作,获得特定格式的数据集,并构建项目特征隶属矩阵和用户项目评分

矩阵。

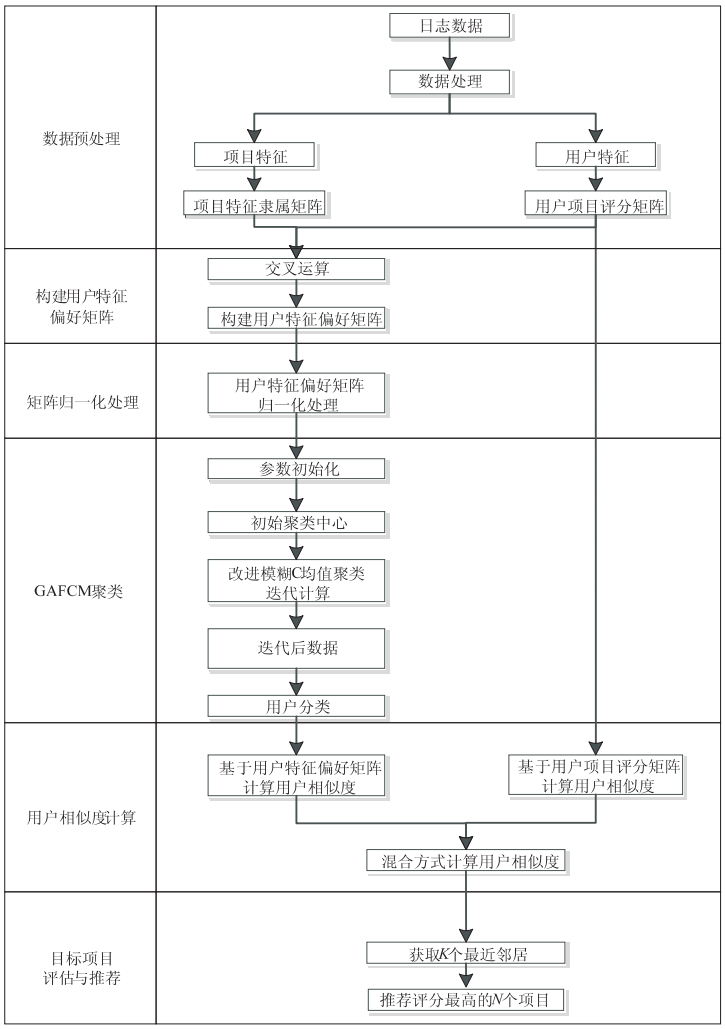

图 1　改进 FCM 的协同过滤流程

3.2　构建用户特征偏好矩阵

时间复杂度、空间复杂度高以及评分矩阵稀疏问题是协同过滤算法目前所面临的主要问题。为了解决用户评分矩阵的稀疏性问题,GAFCM-CF 算法通过利用用户项目评分矩阵和项目特征隶属矩阵来构建用户特征偏好矩阵,构建方法如图 2 所示。

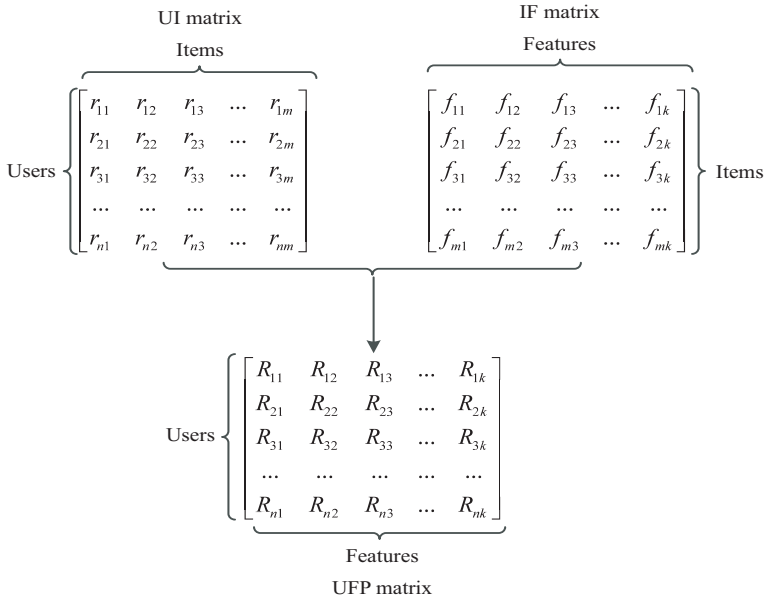

图 2　用户偏好特征矩阵构建过程

图 2 中,矩阵 $\boldsymbol{UI}^{n\times m}$ 为用户项目评分矩阵,矩阵 $\boldsymbol{IF}^{m\times k}$ 为项目特征隶属矩阵,矩阵 $\boldsymbol{UFP}^{n\times k}$ 为用户特征偏好矩阵。可以通过用户项目评分矩阵和项目特征隶属矩阵聚合来构建用户特征偏好矩阵。项目特征隶属矩阵 $\boldsymbol{IF}^{m\times k}$ 中的元素取值为 0 或 1,满足公式(5):

$$f_{i,j}=\begin{cases}0, & \text{项目 } i \text{ 不具备特征 } j\\ 1, & \text{项目 } i \text{ 拥有特征 } j\end{cases} \tag{5}$$

用户 u 对项目的评分向量为 $r_u=(r_{u,1},r_{u,2},\cdots,r_{u,m})$,项目 i 对应特征的隶属向量为 $f_i=(f_{1,i},f_{2,i},\cdots,f_{m,i})$,$R_{ui}$ 计算过程如式(6)所示:

$$R_{ui}=\frac{r_u\times f_i^{\mathrm{T}}}{\sum_{j=0}^{j=m}f_{j,i}} \tag{6}$$

该方法中用户项目评分矩阵通常都是稀疏矩阵,这是由于用户数量和项目数量极多,而单个用户关联的项目数量极少。项目特征隶属矩阵中 k 的取值通常远小于用户评分矩阵中项目的数量 m,因此通过该方法获得的用户对项目特征的偏好矩阵相对于用户项目评分矩阵维度得到了极大降低,有利于降低推荐算法的时间和空间复杂度。

3.3 归一化处理

对 $\boldsymbol{UFP}$ 矩阵进行 min-max 归一化处理,将矩阵各元素数值映射到区间[0,1],映射公式如下所示:

$$x_{i,j}=\frac{x_{i,j}-x_{\min}}{x_{\max}-x_{\min}} \tag{7}$$

其中,$x_{i,j}$ 为矩阵第 i 行第 j 列对应的元素值,在 $\boldsymbol{UFP}$ 矩阵中表示用户 i 对项目特征 j 的偏爱程度,$x_{\min}$ 为所有用户对项目特征偏爱程度中的最小值,$x_{\max}$ 为所有用户对项目特征偏爱程度的最大值。

3.4 GAFCM 聚类

GAFCM-CF 算法为了达到快速收敛并避免局部最优,将遗传算法与 FCM 的算法融合,通过 FCM 算法使数据快速高效地趋于各自的极值点,又可以通过遗传算法摆脱数据在收敛过程中可能陷入的局部最小值的问题[16]。

GAFCM 聚类的具体步骤如下:

步骤 1:对原始数据进行预处理,构建用户偏好特征矩阵 $\boldsymbol{UFP}$ 并对其进行归一化处理。

步骤 2:参数初始化,初始化 GAFCM 算法的相关参数,包括种群大小 M,交叉概率 P_c,变异概率 P_m,最大迭代次数 $t_{\max}$,聚类簇数 c,隶属度因子 m,收敛精度 ε。

步骤 3:编码及种群初始化,根据公式进行编码,并随机产生一个种群 X,X 中有 n 个研究对象作为初始个体,即 $X=[x_1,x_2,\cdots,x_n]$。

步骤 4:计算个体适应度:

$$f_m=\frac{1}{1+\sum_{i=1}^{C}\sum_{j=1}^{n}u_{i,j}^{m}\parallel X_j-c_i\parallel^2} \tag{8}$$

步骤 5:对当前种群执行选择、交叉和变异操作,产生新一代个体。

步骤 6:若 $t=t^{\max}$,遗传算法结束,输出最终的数据,并转入步骤 7;否则,令 $t=t+1$,并返回步骤 4。

步骤 7:根据全局最优解模糊划分整个数据集,输出聚类中心矩阵,实现用户聚类划分。

3.5 用户相似度计算

为计算用户的相似度,GAFCM-CF 算法通过综合利用用户特征偏好矩阵以及用户项目评分矩阵来实现,既包含原始用户项目评分矩阵的显性信息,又考虑到用户对项目特征偏好的隐性信息,如公式(9)所示:

$$\mathrm{Sim}(u,v)=\lambda\ \mathrm{Sim}_1(u,v)+(1-\lambda)\ \mathrm{Sim}_2(u,v) \tag{9}$$

其中,λ 是权重因子,取值范围为(0,1);$\mathrm{Sim}(u,v)$ 表示用户 u 和用户 v 的综合相似度;$\mathrm{Sim}_1(u,v)$ 表示通过公式(1)计算得到的相似度,是使用原始用户项目评分矩阵得到的;$\mathrm{Sim}_2(u,v)$ 表示使用用户对项目特征偏好矩阵得到的相似度,可以通过公式(10)获得:

$$\mathrm{Sim}(u,v)_2=\frac{\sum_{i\in F_{u,v}}(R_{u,i}-\overline{R_u})(R_{v,i}-\overline{R_v})}{\sqrt{\sum_{i\in I_{u,v}}(R_{u,i}-\overline{R_u})^2}\sqrt{\sum_{i\in I_{u,v}}(R_{v,i}-\overline{R_v})^2}} \tag{10}$$

其中,$F_{u,v}$ 表示用户 u 和用户 v 共同偏好的特征的集合;$R_{u,i}$ 是用户 u 对特征 i 的偏好程度;$R_{v,i}$ 是用户 v 对特征 i 的偏好程度;$\overline{R_u}$ 表示用户 u 对所有特征偏好程度的平均值;$\overline{R_v}$ 表示用户 v 对所有特征偏好程度的平均值。

3.6 目标项目评估

用户 u 对项目 i 的评分计算公式为:

$$p(u,i)=\overline{r_u}+\frac{\sum_{v\in S(u)\cup i\in I_{u,v}}\mathrm{Sim}_{u,v}\times(r_{vi}-\overline{r_v})}{\sum_{v\in S(u)}\mathrm{Sim}_{u,v}} \tag{11}$$

其中,S_u 表示与用户 u 相似度最高的前 k 个用户集合;$I_{u,v}$ 表示用户 u 和用户 v 共同评分的项目集合;$\mathrm{Sim}_{u,v}$ 是用户 u 和用户 v 的相似度,由公式(9)获得;r_{vi} 表示用户 v 对项目 i 的评分;$\overline{r_u}$ 表示用户 u 对项目的平均评分;$\overline{r_v}$ 表示用户 v 对项目的平均评分。

4 实验分析

4.1 数据集描述

该文采用 MovieLens 100k 数据集验证算法的性能。该数据集包括 1 682 部电影中的 943 位用户的

100 000 个评分,数据集稀疏度为 93.7%(用户未评分数量占用户最大评分数量的比例)。用户对电影的评分区间为 1 ~ 5 分,每个用户至少评分 20 部电影,用户对某电影的评分值越高表明用户对该电影喜爱程度越大。

该文将原始数据集随机划分为 5 部分,使用 5 折交叉验证方式,每次将其中 4 部分用于训练,剩下的 1 部分用于测试,将 5 次实验的平均值作为实验结果。

4.2　实验设置及评价指标

该文主要通过平均绝对误差(mean absolute error, MAE)、准确率(Precision)和召回率(Recall)三个指标对算法的性能进行分析。

MAE 是衡量预测评分的准确性的重要指标,通过比较预测评分和真实评分之间的平均绝对误差计算得出。MAE 值越小,则表示预测评分与真实评分越接近,算法精度也就越高。Precision 表示正样本在预测为正的样本中所占的比例,即用户发生行为项目占推荐项目的比例。Recall 表示预测为正样本占正样本的比例,即推荐项目占用户产生行为项目的比例。显然,Precision 和 Recall 越大,说明算法的推荐精度越高。

MAE 可以通过公式(12)进行计算:

$$\mathrm{MAE}_u = \frac{\sum_{i=1}^{n} |p_{u,i} - r_{u,i}|}{n} \tag{12}$$

其中,$p_{u,i}$ 表示用户 u 对项目 i 的预测评分,$r_{u,i}$ 表示用户 u 对项目 i 的真实评分,n 表示用户 u 所评分的项目的数量。

Precision 可以通过公式(13)进行计算:

$$\mathrm{Precision} = \frac{\sum_{u \in U} | R(u) \cap T(u) |}{\sum_{u \in U} | R(u) |} \tag{13}$$

Recall 可以通过公式(14)进行计算:

$$\mathrm{Recall} = \frac{\sum_{u \in U} | R(u) \cap T(u) |}{\sum_{u \in U} T | (u) |} \tag{14}$$

上述公式(13)和公式(14)中,U 表示所有项目的集合,$R(u)$ 表示给用户 u 推荐的项目集合,$T(u)$ 表示用户 u 发生行为的项目的集合。

实验相关参数设置如下:模糊聚类分类数 $c = 8$,隶属度因子 $m = 2$,迭代次数 $t = 50$,交叉概率 $P_c = 0.6$,变异概率 $P_m = 0.1$,收敛精度 $\varepsilon = 0.000\ 1$。

4.3　实验结果与分析

首先,对 GAFCM-CF 算法性能随权重因子 λ 的变化情况进行了分析。该组实验将相似用户数量 k 值设置为 20,如图 3 所示,在相似用户数量 $k = 15$ 时,随着 λ 取值逐渐增大,准确率和召回率变化趋势均为先增大后减小,并且在 $\lambda = 0.4$ 时,准确率和召回率达到峰值,分别为 0.251 和 0.129。由图 4 可以看出,随着 λ 取值逐渐增大,平均绝对误差 MAE 变化趋势为先减小后增大,并且在 $\lambda = 0.4$ 时,平均绝对误差取得最小值 0.466。

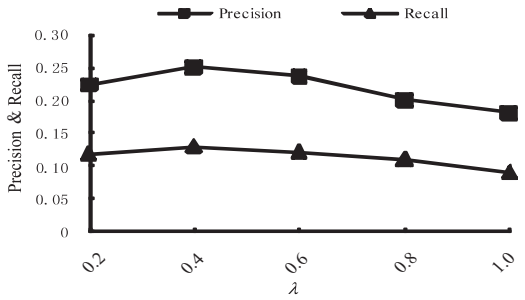

图 3　λ 取值对 Precision 和 Recall 的影响分析

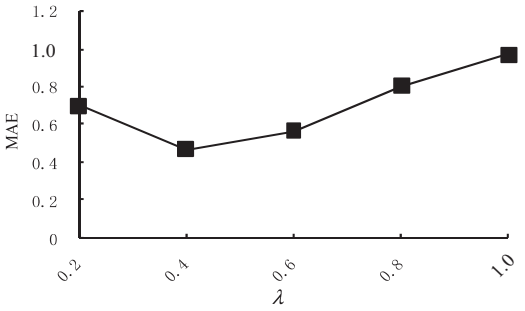

图 4　λ 取值对 MAE 的影响分析

其次,将 GAFCM-CF 算法与文献[6]提出的 PDSFCM 算法、User-CF 算法的进行性能对比,分析了三种算法的 MAE、Precision 和 Recall 随相似用户数量 k 的变化情况。该组实验权重因子 λ 取值均设置为 0.4。

由图 5 可以看出,GAFCM-CF 算法、PDSFCM 算法和 User-CF 算法的 MAE 均随着相似用户数量 k 的增大而减小。在 k 值相同的情况下,GAFCM-CF 算法的 MAE 均比 PDSFCM 算法与 User-CF 算法的 MAE 要小,表明 GAFCM-CF 算法比 User-CF 算法和 PDSFCM 算法具有更好的精度。

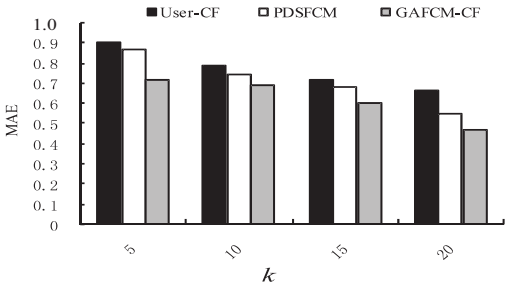

图 5　MAE 对比分析

由图 6 和图 7 可以看出,GAFCM-CF 算法、PDSFCM 算法及 User-CF 算法的 Precision 和 Recall 均随着相似用户数量 k 的增大而增大。在 k 值相同的情况下,GAFCM-CF 算法的预测准确率和召回率都比

User-CF 算法和 PDSFCM 算法的预测准确率和召回率要高,表明 GAFCM-CF 算法比 User-CF 算法和 PDSFCM 算法具有更好的推荐效果。

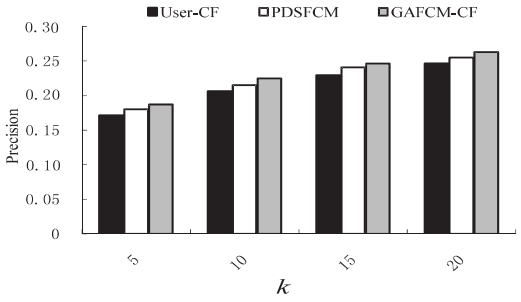

图 6 Precision 对比分析

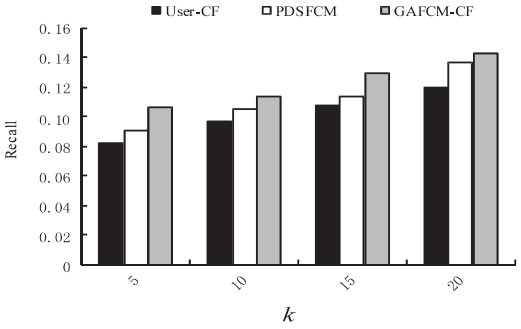

图 7 Recall 对比分析

5 结束语

针对传统协同过滤推荐算法中存在的数据稀疏性及推荐准确率低的问题,提出了一种基于改进 FCM 的协同过滤推荐算法 GAFCM-CF。实验结果表明,相比于传统的基于用户的协同过滤推荐算法,该算法具有更高的推荐质量以及推荐准确率。未来工作中,将考虑进一步挖掘用户隐藏信息,进一步提升推荐算法的准确率;另一方面,将对算法的复杂度和其他方面的推荐性能,比如推荐物品的覆盖率、流行度、惊喜度等进行更全面的评估。

参考文献:

[1] 刘华锋,景丽萍,于 剑. 融合社交信息的矩阵分解推荐方法研究综述[J]. 软件学报,2018,29(2):340-362.

[2] 项 亮. 推荐系统实践[M]. 第 3 版. 北京:人民邮电出版社,2012.

[3] SANDVIG J,MOBASHER BURKE R. A survey of collaborative recommendation and the robustness of model-based algorithms[J]. IEEE Data Eng. Bull,2008,31:3-13.

[4] KATARYA R,VERMA O P. An effective web page recommender system with fuzzy c-mean clustering[J]. Multimedia Tools and Applications,2017,76:21481-21496.

[5] THONG N T,SON L H. HIFCF:an effective hybrid model between picture fuzzy clustering and intuitionistic fuzzy recommender systems for medical diagnosis[J]. Expert Systems with Applications,2015,42(7):3682-3701.

[6] MA X,LU H,GAN Z,et al. Improving recommendation accuracy with clustering-based social regularization[C]//Web technologies and applications. Changsha, China: Springer, 2014:177-188.

[7] TRAN C,KIM J,SHIN W,et al. Clustering-based collaborative filtering using an incentivized/penalized user model[J]. IEEE Access,2019,7:62115-62125.

[8] MIRBAKHSH N,LING C X. Improving Top-N recommendation for cold-start users via cross-domain information[J]. ACM Transactions on Knowledge Discovery from Data, 2015,9(4):1-19.

[9] WANG J,ZHANG N,YIN J. Collaborative filtering recommendation based on fuzzy clustering of user preferences [C]//2010 seventh international conference on fuzzy systems and knowledge discovery. Yantai, China: IEEE, 2010: 1946-1950.

[10] YING Y, CAO Y. Collaborative filtering recommendation combining FCM and slope one algorithm[C]//2015 international conference on informative and cybernetics for computational social systems (ICCSS). Chengdu, China: IEEE, 2015:110-115.

[11] LEMIRE D,MACLACHLAN A. Slope one predictors for online rating-based collaborative filtering[C]//SIAM data mining (SDM'05). Newport Beach, California:[s. n.], 2005:21-23.

[12] 朱敬华,王 超,马胜超. 基于社交信任聚类的混合推荐算法[J]. 软件学报,2018,29(S1):21-31.

[13] IFADA N,PRASETYO E H,AND MULA'AB. Employing sparsity removal approach and fuzzy c-means clustering technique on a movie recommendation system[C]//2018 international conference on computer engineering,network and intelligent multimedia (CENIM). Surabaya, Indonesia: IEEE,2018:329-334.

[14] AL-BAKRIN F,HASSAN S. A proposed model to solve cold start problem using fuzzy user-based clustering[C]// 2019 2nd scientific conference of computer sciences (SCCS). Baghdad,Iraq:[s. n.],2019:121-125.

[15] PANDYA S,SHAH J,JOSHI N,et al. A novel hybrid based recommendation system based on clustering and association mining[C]//10th international conference on sensing technology. Nanjing,China:[s. n.],2016:1-6.

[16] TANG Hailiang,SHI Lei,LIU Bin. The application research of FCM clustering based on genetic algorithm in the telephone user's behavior[C]//Proceedings of the 2016 2nd workshop on advanced research and technology in industry applications. Dalian,China:[s. n.],2016:1725-1731.