融合 FCM 和 TFNs 的协同过滤推荐算法

徐新卫[1],陶 飞[1],邓佳佳[1],周 俊[2]
(1. 安徽工业大学 管理科学与工程学院,安徽 马鞍山 243032;
2. 广东科技学院,广东 东莞 523073)

摘 要:针对推荐算法中的稀疏性问题和传统推荐系统中使用离散评分,用户对物品的喜好程度只能通过 5 个等级来选取,用户对物品的偏好程度是模糊的且 5 等级评分不能合理表达用户的喜好,提出一种结合模糊 C 均值(Fuzzy C-Means,FCM)和梯形模糊数(Trapezoidal Fuzzy Numbers,TFNs)的协同过滤算法。首先,在传统的模糊 C 均值算法上融合遗传算法,将遗传算法的搜索结果作为模糊 C 均值的初始聚类中心,以其克服传统 FCM 搜索极易陷入局部最小值点的缺陷;然后,引入梯形模糊相似度模型,将离散评分数转化为梯形模糊数以此来计算用户相似度,从而利用模糊分数预测估计进行推荐;最后,选取 MAE 和 RMSE 作为评估指标,在 Movielens 数据集中进行实验,实验结果显示所提算法在与其余四种算法对比中预测误差更低,精确度更高,有效提高了推荐质量,也证明了该算法对于稀疏性问题有一定程度上的缓解,表明了该算法的有效性。

关键词:协同过滤;梯形模糊数;模糊 C 均值;遗传算法;离散评分
中图分类号:TP391.3 **文献标识码**:A **文章编号**:1673-629X(2023)03-0161-06
doi:10.3969/j.issn.1673-629X.2023.03.024

Collaborative Filtering Recommendation Algorithm Incorporating FCM and TFNs

XU Xin-wei[1],TAO Fei[1],DENG Jia-jia[1],ZHOU Jun[2]
(1. School of Management Science and Engineering,Anhui University of Technology,Maanshan 243032,China;
2. Guangdong Institute of Science and Technology,Dongguan 523073,China)

Abstract:In view of the sparsity problem in recommendation algorithms and the discrete ratings used in traditional recommendation systems,the user's preference for items only be selected by 5 levels,the fuzzy user's preference for items and the 5-level rating cannot reasonably express the user's preference,a collaborative filtering algorithm combining Fuzzy C-Means (FCM) and Trapezoidal Fuzzy Numbers (TFNs) is proposed. Firstly,the genetic algorithm is integrated with the traditional fuzzy c-means algorithm,and the search result of the genetic algorithm is used as the initial clustering center of the fuzzy c-means to overcome the defect that the traditional FCM search is quite easy to fall into the local minima. Then the trapezoidal fuzzy similarity model is introduced,and the discrete score numbers are transformed into trapezoidal fuzzy numbers to calculate the user similarity,so that the fuzzy score prediction estimation can be used for recommendation. Finally,MAE and RMSE are selected as evaluation indexes and experiments are conducted in Movielens dataset. It is showed that the proposed algorithm has lower prediction error and higher accuracy in comparison with the remaining four algorithms,which effectively improves the recommendation quality and also proves that the algorithm has a certain degree of alleviation for the sparsity problem,indicating its effectiveness.

Key words:collaborative filtering;trapezoidal fuzzy number;fuzzy C-means;genetic algorithm;discrete scoring

0 引 言

伴随信息技术和互联网的高速发展,数据量表现出几何级稳定增长。而推荐系统(Recommendation Systems,RSs)是有效的信息处理手段之一。为了解决信息过载的问题,RSs 主要采用信息过滤的技术,通过用户的历史行为信息提取出用户偏好,减少用户查找信息的时间,并将用户与商品相互联。协同过滤(Collaborative Filtering,CF)是推荐最成功的技术之一,根

收稿日期:2022-05-18 **修回日期**:2022-09-22
基金项目:安徽省省级教学研究项目(2019jyxm0145)
作者简介:徐新卫(1971-),男,副教授,博士,研究方向为智能计算、数据挖掘;通信作者:陶 飞(1996-),男,硕士研究生,研究方向为数据挖掘、机器学习、隐私保护。

据相似度指数发现与当前用户相似的用户,然后利用这些相似用户对产品进行评估。基于用户的 CF 算法(User-Based CF,UBCF)和基于项目的 CF 算法(Item - Based CF, IBCF)是协同过滤推荐算法(Recommendation Algorithms,RAs)的两种基本形式。算法主要是研究用户行为定位相似的用户。IBCF 是利用项目间的相似性来提出与附近用户感兴趣的事物相似的东西。CF 算法可以帮助客户找到合适的物品或服务,但它们被诸如冷启动、数据稀少和可扩展性等问题所困扰。

近年来,学者们利用聚类方法[1]来解决 CF 算法中数据系数以及推荐精度低等问题。贾俊杰等[2]利用了信任的传递性,先计算出显隐式信任获得综合直接信任值,然后再获得 Jaccard 全局信任值,最后将综合直接信任值和 Jaccard 全局信任值结合得到综合信任值,将其融入 FCM 中实现推荐;Tran C 等[3]提出了一种新的基于聚类的 CF 方法,仅仅使用激励/惩罚用户模型(Incentivized/Penalized User,IPU),只由用户给出评级来对用户进行聚类,无需进一步的事先信息从而进行推荐;张建华等[4]针对知识 RAs 的语义缺失和精准性低的问题,提出一种基于改进的 LDA-FCM 的知识 RAs,利用 LDA 模型挖掘用户知识模型并用 FCM 聚类用户,再结合 JS 散度代替欧氏距离以提高推荐精度;王永贵等[5]提出一种优化聚类的 CF 算法,根据评分差异对原始评分矩阵进行预处理得到用户项目评分矩阵及项目类型矩阵,构建用户类型偏好矩阵,缓解了数据的稀疏性;赵学健等[6]利用遗传算法(Genetic Algorithm,GA)改进 FCM,同时考虑用户特征偏好矩阵和用户项目评分矩阵计算相似度实现推荐;郑鑫等[7]采用 FCM 筛选最近邻居,结合奇异值分将用户评分分解为不同特征以实现推荐。

通过上述分析,可以发现利用聚类方法对用户进行聚类时,一方面容易受到异常值和噪声的干扰。另一方面,FCM 算法的结果并不是很稳定,因为它的起始质心是随意创建的。需要注意的是,基于协同过滤推荐算法(Collaborative Filtering Recommendation Algorithms,CFRAs)进行相似性度量是至关重要的。UBCF 和 IBCF 算法中最常用的相似度量是余弦(Cosine)、余弦修正(Adjusted Cosine,AC)和皮尔逊相关系数(Pearson Correlation Coefficient,PCC)。但在用户行为复杂的情况下,PCC、AC 和余弦作为相似度度量的基于 CFRAs 的性能上无法得到保证。因此,学者们在设计更全面的相似度方面做出了巨大的努力,将用户评分的信息量及模糊性考虑其中以计算相似度,提升推荐精度。

例如,王森等[8]利用梯形模糊数(Trapezoidal Fuzzy Numbers,TFNs)表示评分与满意度的映射关系,融合基于模糊评分的项目相似性和基于标签隶属度的相似性形成项目相似度,从而进行推荐;Zhang 等[9]为确定用户的相似度,根据三角模糊数的相似度,将三角形面积和中点用来表达用户对项目的整体评价,从而使相似度计算的准确性得到提高;Yager[10]针对用户分数潜在的信息,采用模糊理论中的模糊子集并结合 PCC 算出评分的相似度,其后对两部分加权获得最终相似度;吴毅涛[11]考虑多数推荐系统使用离散评分,用户只能在评分级别{1,2,3,4,5}中选择相应的分数,提出了梯形模糊评分模型,将评分模糊化并考虑评分信息量和模糊性,用梯形模糊数来计算相似度以提高推荐质量,并证明模糊相似度(Fuzzy Similarity,FS)是余弦在模糊域上的扩展。

因此,该文提出了一种融合 FCM 和 TFNs 的协同过滤推荐算法(Genetic Algorithm Improved Fuzzy C-means And Trapezoidal Fuzzy Number Collaborative Filtering Algorithm,GAFCM-TFNCF)。首先,使 FCM 与 GA 相结合,然后,对用户进行聚类,最后,根据 FS 将离散分数转换成梯形模糊数用来计算用户相似度,从而利用模糊分数预测评分。此方法的优点如下:

(1)将 GA 的搜索结果作为 FCM 的初始聚类中心,可以有效避免 FCM 在迭代过程中陷入局部最小值点;

(2)现有 RSs 下的评分只能片面地表达用户的偏好,该文通过引入 TFNs 将评分模糊化,从而计算用户的相似度,这样更能合理表达用户的观点。

1 相关工作

1.1 FCM 算法

在实际应用中,基于 FCM 方法是使用最广泛的模糊聚类技术。它允许数据项属于隶属度从 0 到 1 不等的集合,通过优化目标函数来定义样本点的类别,以获得每个样本点对所有类中心的隶属度,从而达到样本数据自动分类的目的。基本如公式(1)所示:

$$\min J_{\mathrm{FCM}}(H,V)=\sum_{k=1}^{m}\sum_{i=1}^{c}(\mu_{ik})^{M}\parallel u_k-v_i\parallel^2 \quad (1)$$

其中,m 是给定数据集中的数据向量总数;c 为聚类数目;$V=\{v_1,v_2,\cdots,v_i,\cdots,v_c\}$ 表示聚类中心;$\mu_{ik}(\mu_{ik}\in[0,1])$ 是用户 u_k 对聚类中心 $v_i(i=1,2,\cdots,c)$ 的隶属度;$\boldsymbol{H}_{m\times c}=(\mu_{ik})$ 是一个模糊分区矩阵,也为聚类的结果;M 是 $J_{\mathrm{FCM}}(H,V)$ 中的一个参数,一般 $M\in[1.25,2.5]$,通常称为模糊器,用来控制成员等级影响的实数。

聚类中心 v_i 和隶属度矩阵 $\boldsymbol{H}$ 根据以下的公式来计算,使得 $J_{\mathrm{FCM}}(H,V)$ 最小化。

$$d_{ik} = \| u_k - v_i \|^2 \tag{2}$$

$$\mu_{ik} = \frac{1}{\sum_{j=1}^{c} \left(\frac{d_{ik}}{d_{jk}}\right)^{\frac{2}{M-1}}} \tag{3}$$

$$v_i = \frac{\sum_{k=1}^{m} (\mu_{ik})^M u_k}{\sum_{k=1}^{m} (\mu_{ik})^M} \tag{4}$$

$$\sum_{i=1}^{c} \mu_{ik} = 1 \tag{5}$$

式(2)中,$d_{ik}(u_k, v_i)$ 是数据对象 u_k 和簇 v_i 之间的距离值,欧几里得度量来衡量距离值是常用方法之一。FCM 算法旨在将方程中的准则函数最小化,距离值 d_{ik} 是通过迭代更新式(3)中的聚类中心 v_i 和式(4)中的隶属度 μ_{ik},迭代更新重复过程,直到达到条件即终止。

FCM 算法流程可以如下表示[6]:

(1)输入聚类数目 c,模糊器 M 和距离函数 $\| \cdot \|$;

(2)初始化聚类中心 $v_i (i = 1,2,\cdots,c)$;

(3)根据公式(2)、(3)和(5),计算 $u_{ki} (k = 1,2,\cdots,m; i = 1,2,\cdots,c)$;

(4)利用公式(4)计算聚类中心 $v_i^1 (i = 1,2,\cdots,c)$;

(5)如果满足 $\max\limits_{1 \leqslant i \leqslant c} (\| v_i^0 - v_i^1 \| / \| v_i^1 \|) \leqslant \varepsilon$,则进行下一步;否则,让 $v_i^0 = v_i^1 (i = 1,2,\cdots,c)$ 回到(3)继续执行;

(6)输出聚类中心 $v_i^1 (i = 1,2,\cdots,c)$ 和隶属度矩阵 $\boldsymbol{H}$;

(7)停止。

1.2　模糊集和梯形模糊数的运算

在推荐系统中,用户需要从给定的评分集合{1,2,3,4,5}中选择一个评分,但是有限的评分不能合理表达用户对项目的偏好程度,往往选择一个接近的评分来表达偏好程度[8]。因此,本节提出利用梯形模糊数模型来表示用户评分之间的关系。Zadeh[12]在 1968 年提出了模糊理论,对模糊子集的定义为:给定论域 U,U 中的模糊集 $\tilde{A}$ 由特征函数 $\mu_{\tilde{A}}$ 来表示,如式(6)所示。

$$\mu_{\tilde{A}}: U \to [0,1] \tag{6}$$

模糊集是一类具有连续成员级别的项目集合,由一个隶属函数 $\mu_{\tilde{A}}$ 定义,该函数为每个对象分配一个 0 到 1 之间的隶属度。根据 Ahmad S[13]给出的定义,梯形模糊数为 $\tilde{A} = (a_1, a_2, a_3, a_4; w_A)$,$\tilde{A}$ 是 $\mathbb{R}$ 上的模糊集,a_1、a_2、a_3、a_4 表示梯形模糊数 $\tilde{A}$ 的各个顶点的横坐标值,$a_1 \leqslant a_2 \leqslant a_3 \leqslant a_4$,$w_A$ 为模糊数的最大隶属度,$0 < w_A \leqslant 1$,隶属函数如式(7)所示:

$$\mu_A(x) = \begin{cases} w_A \dfrac{x - a_1}{a_2 - a_1}, & x \in [a_1, a_2) \\ w_A, & x \in [a_2, a_3] \\ w_A \dfrac{a_4 - x}{a_4 - a_3}, & x \in (a_3, a_4] \\ 0, & \text{其它} \end{cases} \tag{7}$$

假设存在 $\tilde{b} = (b_1, b_2, b_3, b_4)$ 和 $\tilde{c} = (c_1, c_2, c_3, c_4)$ 两个梯形模糊数,这两个模糊数定义如下式所示:

$$\tilde{b} \oplus \tilde{c} = (b_1 + c_1, b_2 + c_2, b_3 + c_3, b_4 + c_4) \tag{8}$$

$$\tilde{b} \ominus \tilde{c} = (b_1 - c_4, b_2 - c_3, b_3 - c_2, b_4 - c_1) \tag{9}$$

$$\tilde{b} \otimes \tilde{c} = (b_1 * c_1, b_2 * c_2, b_3 * c_3, b_4 * c_4) \tag{10}$$

$$\tilde{b} \oslash \tilde{c} = (b_1 / c_4, b_2 / c_3, b_3 / c_2, b_4 / c_1) \tag{11}$$

该文将利用上述公式改进目前 CFRAs 中模糊相似度的计算,从而进行评分预测。

2　基于改进 FCM 和 TFNs 的协同过滤算法

针对模糊相似度的计算,目前 CF 算法仅采用常规距离和重心距离,存在一定的误差。因此,该文考虑 Ahmad S[13]提出的一种新的梯形相似性度量,它结合了几何距离(Geometric Distance, GD)、重心距离(Center Of Gravity, COG)、Dice 相似性系数(Dice Similarity Coefficient, DSC)和 Hausdorff 距离(Hausdorff Distance, HD)。CF 算法常用的是 PCC、Cosine 等计算相似度,但未能考虑评分信息量,并不适用于模糊评分的相似度计算,所以本节采用梯形相似度结合评分信息来进行推荐。

2.1　梯形模糊相似度

给定一个连续的全域 U 和在 U 上的一组模糊数 FS(U)[14],假设两个梯形模糊评分为 $\tilde{A}_i = (a_1, a_2, a_3, a_4; w_{\tilde{A}})$ 和 $\tilde{B}_j = (b_1, b_2, b_3, b_4; w_{\tilde{B}})$,根据 Ahmad[13,15]的定义,相似度计算公式如式(12)所示:

$$S_1(\tilde{A}_i, \tilde{B}_j) = \left(1 - \frac{1}{4}\sum_{k=1}^{4} |a_i - b_j|\right) \times \left(1 - |\hat{x}_{\tilde{A}} - \hat{x}_{\tilde{B}}|\right)^{B(S_1, S_2)} \times \left(\frac{2W_{\tilde{A}}W_{\tilde{B}}[(a_1 + a_2)(b_1 + b_2) + (a_3 + a_4)(b_3 + b_4)]}{(w_{\tilde{A}}^2)[(a_1 + a_2)^2 + (a_3 + a_4)^2] + (w_{\tilde{B}}^2)[(b_1 + b_2)^2 + (b_3 + b_4)^2]}\right) \times \left(\frac{1}{1 + [max\{|a_1 - b_1|, |a_2 - b_2|, |a_3 - b_3|, |a_4 - b_4|\} + |W_{\tilde{A}} - W_{\tilde{B}}|]}\right) \tag{12}$$

其中,$\hat{x}_{\tilde{A}}$ 和 $\hat{y}_{\tilde{A}}$ 是梯形 $\tilde{A}$ 和 $\tilde{B}$ 的水平重心,如式(13)、(14)所示。几何距离(Geometric Distance, GD)、重心距离(Center Of Gravity, COG)、Dice 相似性系数(Dice Similarity Coefficient, DSC)和 Hausdorff 距离

(Hausdorff Distance, HD)是式(12)中的变量[8]。GD 衡量了两个梯形的横向距离,梯形重心的水平与垂直距离为 COG,重心纵坐标与梯形上下底的长度差相关联,所以 COG 反映了模糊数信息量的差异,信息差越大,COG 越大,为判断两个集合的相似性采用 DSC,它表明两个集合的重复率。$w_{\tilde{A}}$、$w_{\tilde{B}}$ 表示梯形 $\tilde{A}$、$\tilde{B}$ 的信息量,$B(S_{\tilde{A}},S_{\tilde{B}})$ 决定是否使用 COG,就是当 $a_4-a_1+b_4-b_1=0$ 时,不使用 COG,定义如式(15)所示。

$$\hat{x}_{\tilde{A}}=\frac{\hat{y}_{\tilde{A}}(a_2+a_3)+(w_{\tilde{A}}-\hat{y}_{\tilde{A}})(a_1+a_4)}{2w_{\tilde{A}}} \tag{13}$$

$$\hat{y}_{\tilde{A}}=\begin{cases}\dfrac{w_{\tilde{A}}}{6}\left(\dfrac{a_3-a_2}{a_4-a_1}+2\right) & ,a_1\neq a_4,\ 0<w_{\tilde{A}}\leqslant 1\\ \dfrac{w_{\tilde{A}}}{2} & ,a_1=a_4,\ 0<w_{\tilde{A}}\leqslant 1\end{cases} \tag{14}$$

$$B(S_{\tilde{A}},S_{\tilde{B}})=\begin{cases}1, & a_4-a_1+b_4-b_1>0\\ 0, & a_4-a_1+b_4-b_1=0\end{cases} \tag{15}$$

根据 GD 和 COG 可以看出两个用户对于一个项目的相似度,再通过加权所有项目相似度获得用户模糊相似度,公式如下:

$$\mathrm{sim}(u,v)=\frac{n(U)}{n}-\frac{\sum_{i\in U}|R_{u,i}-R_{v,i}|}{n} \tag{16}$$

由用户 u 和 v 共同评价的项目集表示为 U,数量表示为 $n(U)$,项目 i 被用户 u 评价的分数为 $R_{u,i}$,而用户 u 评价的项目数是 n。

2.2 GAFCM-TFNCF 算法流程

因为遗传算法[16-17]在生成优化和搜索问题方面有着较不错的解决方案,目前许多学者已将 GA 用于处理聚类问题或者将多种不同的方式用来集成 GA 和聚类方法[18-20],且使用 GA 优化初始聚类中心和聚类方法的参数[21-22]。所以在用 FCM 算法进行聚类之前,先利用遗传算法对其进行改进。所提出的模型步骤如下[23]:

(1)对原始数据集进行预处理,搭建出用户偏好矩阵,并将其进行归一化处理;

(2)将 GAFCM 算法的参数 M(种群规模)、m(隶属度因子)、c(聚类的簇数)、p_c(交叉率)、p_m(变异率)、$t_{\max}$(最大迭代次数)、ε(收敛的精度)进行初始化;

(3)将从数据集中随机选择初始染色体;

(4)更新最佳染色体;

(5)选择 t 染色体,利用轮盘赌选择放入选择池中,其轮盘赌是 GA 的一种选择机制,公式如下所示:

$$P_i=\frac{f_i}{\sum_{j=1}^{t}f_j} \tag{17}$$

其中,P_i 为染色体 $i(1\leqslant i\leqslant t)$ 的选择概率,f_i 和 f_j 分别是 i 和 j 的适应值。

(6)随机将两条染色体作为双亲并检查交叉概率,以用来核实是否进行交叉步骤,在交叉过程中生成新染色体的概率定义如下:

$$X_g^{\mathrm{new}}=rX_g+(1-r)Y_g \tag{18}$$

$$Y_g^{\mathrm{new}}=rY_g+(1-r)X_g \tag{19}$$

其中,X_g^{new} 和 Y_g^{new} 是在交叉后后代的基因,X_g 和 Y_g 为父母基因,r 是 0 到 1 之间的随机数,通常交叉概率设置在[0.85,0.95]之间。

(7)核实突变 p_m 的概率,以确定是否继续突变步骤,实施编码遗传突变概率定义如下:

$$X_g^{\mathrm{new}}=X_g+s_g\cdot r_g\cdot a_g \tag{20}$$

$$a_g=2^{-u\cdot k} \tag{21}$$

其中,s_g 和 u 为随机数,且 $s_g\in\{-1,1\}$ 和 $u\in[0,1]$,r_g 为变异的范围,$r_g\in[10^{-6},0.1]$;$k\in\{4,5,\cdots,20\}$ 表示变异的精度。

(8)若满足终止条件则结束,否则转到步骤(3);

(9)将 GA 的输出结果作为 FCM 的初始聚类中心;

(10)利用(2)~(5)进行计算获得聚类中心和隶属度矩阵,从而实现用户聚类划分。

在经过上述操作后,利用梯形模糊评分来计算相似度并进行推荐。

(1)相似度计算流程如下:

输入:用户项目评分矩阵 $\boldsymbol{R}$,邻居数量 k。

①根据(12)计算目标用户 u 和其他用户 v 在一个任务的共同评价方面的相似度 $S(R_{u,i},R_{v,i})$;

②利用式(16)计算模糊相似度 $\mathrm{sim}(u,v)$。

(2)产生推荐集。

①作为邻居集,选择相似度最高的 k 个用户;

②如式(22)所示,采用平均加权方法对预测评分进行加权。

$$P_{u,i}=\bar{R}_u p+\frac{\sum_{v\in K}\mathrm{sim}(u,v)(R_{v,i}-\bar{R}_v p)}{\sum\mathrm{sim}(u,v)} \tag{22}$$

其中,用户 u 对项目 i 的预测得分是 $P_{u,i}$,用户 v 任务得分均值是 $\bar{R}_v$,用户 u 和用户 v 之间的相似度为 $\mathrm{sim}(u,v)$,此外还通过 p 限制平均值对预测评分的影响程度。

输出:目标用户 u 对未知任务的预测评分,获得推荐结果。

3 实验结果分析

3.1 实验环境及实验数据

实验的操作系统为 Windows10,CPU 为 Intel(R)

Core(TM) i5-1035G1@ 1.00 GHz 1.19 GHz,实验内存为 16 GB,主要实验平台是 Python 3.7。为验证 GAFCM-TFNCF 算法的性能,利用常见的数据集 Movielens 进行实验,其数据集的描述如表 1 所示。

表 1　实验数据集

数据集	用户数	项目数	评分数	稀疏度/%
100 K	943	1 682	100 000	93.7
1 M	6 040	3 952	1 000 000	95.81

3.2　评价指标实验数据设置

将平均绝对误差(Mean Absolute Error,MAE)、均方根误差(Root Mean Squared Error,RMSE)作为实验的度量标准;MAE 和 RMSE 的值越小,预测的估计值就越接近真实值,推荐的准确性就越好。具体定义如公式(23)、(24)所示。

$$\mathrm{MAE}=\frac{\sum_{u,i\in T}|P_{u,i}-R_{u,i}|}{|T|} \tag{23}$$

$$\mathrm{RMSE}=\sqrt{\frac{1}{T}\sum_{u,i\in T}(P_{u,i}-R_{u,i})^2} \tag{24}$$

其中,用户 u 对项目 i 的预测评分值为 $P_{u,i}$;用户 u 对项目 i 的真实评分值是 $R_{u,i}$;T 为测试集;$|T|$ 为测试集的数量。

相关参数设置如下:种群规模 $M=50$,迭代数 $t=50$,变异概率 $p_m=0.05$,交叉概率 $p_c=0.85$,收敛的精度 $\varepsilon=0.000\ 1$,隶属度因子 $m=2$,聚类的簇数 $c=8$。

3.3　算法对比

为了验证算法 GAFCM-TFNCF 的有效性,对比算法有 UBCF、基于用户偏好和项目属性的协同过滤推荐算法 UPPPCF[24]、传统基于 FCM 的协同过滤算法 FCMCF[7]、基于项目模糊相似度的协同过滤推荐算法 IFSCF[8]。

3.4　实验结果

将 80% 的数据集作为训练集,剩下的 20% 作为测试集。对比算法的参数均根据自身文献设置为最优并记录下结果以进行对比。使用 MAE、RMSE 来衡量预测评分的准确性。为减小由于分割数据集所带来的误差,将 10 次实验的平均值作为结果。以邻居数量当作变量,间隔为 5,比较文中算法(GAFCM-TFNCF)与其它四种算法的预测精度,实验结果如图 1 所示。

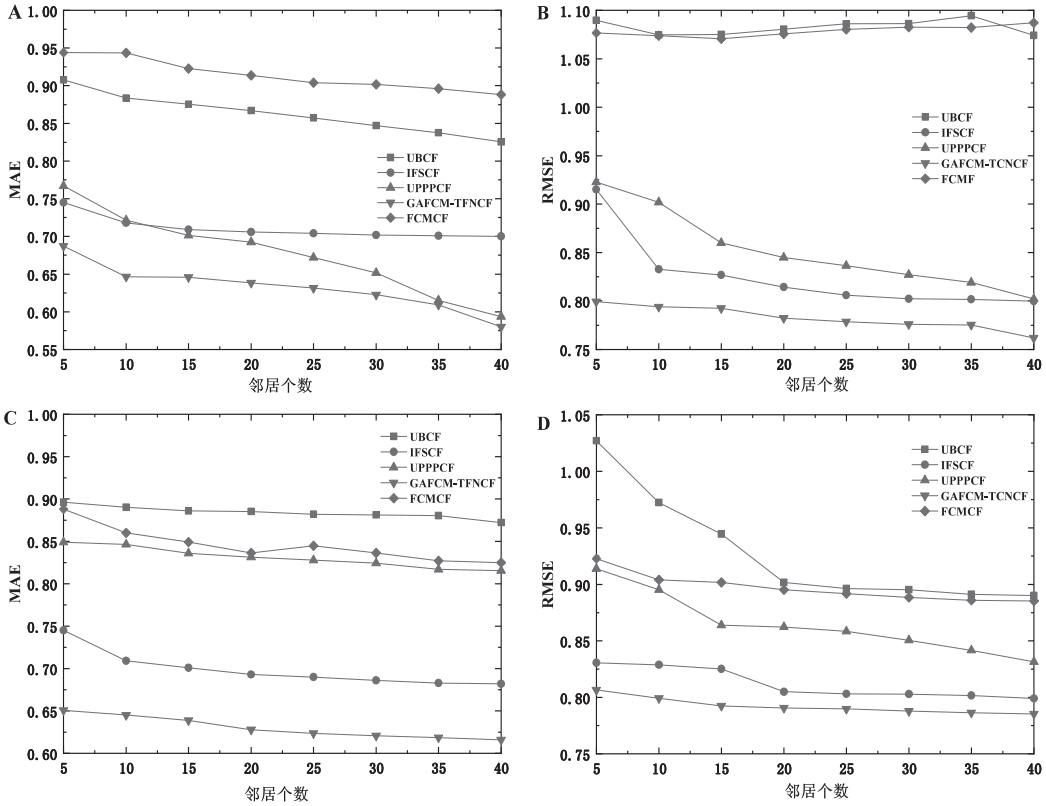

图 1　五种算法在数据集 100 K 和 1 M 上的 MAE 和 RMSE 对比

由图 1 可以看出,GAFCM-TFNCF 算法在所有的对比算法中表现最佳。GAFCM-TFNCF 算法相比其余四种算法 MAE 值要小很多。这表明融合 GA 和 FCM 进行聚类,极大避免了在搜索过程中陷入极小值点,然后再将用户评分模糊化能够极大地提高推荐的精准度。对比 UPPPCF 算法,当邻居数量不断变大后,因为用户对相同项目的评分变得很少,但 GAFCM-TFNCF 算法的 MAE 曲线相较于 UPPPCF 算法更平稳,这表明文中算法在性能表现上更加稳定,不易受到数据集大小的影响,而 UPPPCF 算法下降幅度较大,说

明容易受到数据集大小的影响。

为了比较数据稀疏性对算法精度的影响,实验在 Movielens-100 K 数据集中进行,核定用户数和项目数不变的前提下,将评分数逐步进行减少,比较五种算法的 MAE 值,实验结果如图 2 所示。

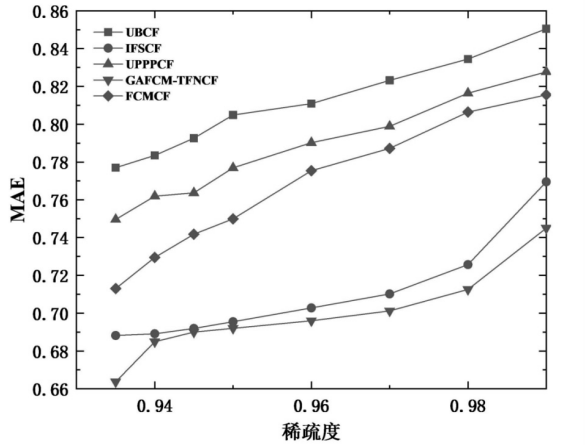

图 2 不同数据稀疏度下的五种算法对比

通过图 2 发现,随着稀疏度的不断增加,可使用的数据不断减少,五种算法的 MAE 值都不断增长,稀疏度在大于 95% 后增长速度变快。GAFCM-TFNCF 算法增长幅度较其余四种算法小,推荐精度更高,表明该算法可以在数据稀疏中完成推荐,其中 UBCF 算法预测最差,这表明该算法对于稀疏度高的数据集并不适用。

4 结束语

针对推荐系统存在数据稀疏性、未考虑用户评分信息量等问题,提出一种基于改进 FCM 的用户模糊相似度的协同过滤推荐算法。对于传统 FCM 聚类对异常值和噪声很敏感的问题,采用 GA 融合 FCM,避免 FCM 在开始搜索时陷入局部最小值点。同时兼顾梯形模糊评分模型,将离散评分模糊化用来计算相似度,从而更合理地表达了用户对项目的偏好。在推荐算法中采用常见数据集进行实验对比,表明算法具有更优的推荐精准度。在未来工作中,考虑在保证推荐准确度的同时,如何同时保护用户的隐私,这是今后的一个研究方向。

参考文献:

[1] KHOSHNESHIN M, STREET W N. Incremental collaborative filtering via evolutionary co-clustering[C]//ACM conference on recommender systems. Barcelona: DBLP, 2010: 325.

[2] 贾俊杰,张玉超. 基于用户模糊聚类的综合信任推荐算法[J]. 计算机工程,2021,47(6):60-67.

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

[4] 张建华,冉 佳,刘 柯. 基于改进 LDA-FCM 的 UserCF 知识推荐研究[J]. 科技管理研究,2020,40(19):140-146.

[5] 王永贵,刘凯奇. 一种优化聚类的协同过滤推荐算法[J]. 计算机工程与应用,2020,56(15):66-73.

[6] 赵学健,张雨豪,陈 昊,等. 基于 FCM 用户聚类的协同过滤推荐算法[J]. 计算机技术与发展,2021,31(8):6-12.

[7] 郑 鑫,张韧志. 一种基于模糊 C 均值聚类的协同过滤推荐算法[J]. 济南大学学报:自然科学版,2016,30(1):55-59.

[8] 王 森,陈 莉,张 洁. 基于项目模糊相似度的协同过滤推荐算法[J]. 计算机应用研究,2021,38(3):696-701.

[9] ZHANG X, MA W, CHEN L. New similarity of triangular fuzzy number and its application[J]. The Scientific World Journal, 2014, 2014: 215047.

[10] YAGER R R. Fuzzy logic methods in recommender systems[J]. Fuzzy Sets and Systems, 2003, 136(2): 133-149.

[11] 吴毅涛,张兴明,王兴茂,等. 基于用户模糊相似度的协同过滤算法[J]. 通信学报,2016,37(1):198-206.

[12] ZADEH L A. Probability measures of fuzzy events[J]. Journal of Mathematical Analysis and Applications, 1968, 23(2): 421-427.

[13] AHMAD S A S, MOHAMAD D, SULAIMAN N H, et al. A distance and set theoretic-based similarity measure for generalized trapezoidal fuzzy numbers[C]//Proceeding of the 25th national symposium on mathematical sciences (SKSM25): mathematical sciences as the core of intellectual excellence. Pahang: [s. n.], 2018.

[14] GUHA D, CHAKRABORTY D. A new approach to fuzzy distance measure and similarity measure between two generalized fuzzy numbers[J]. Applied Soft Computing, 2010, 10(1): 90-99.

[15] GOLDBERG D E. Genetic algorithms in search, optimization, and machine learning[J]. Ethnographic Praxis in Industry Conference Proceedings, 1989, 9(2): 80-85.

[16] PIZZUTI C, PROCOPIO N. A K-means based genetic algorithm for data clustering[C]//International joint conference SOCO'16-CISIS'16-ICEUTE'16. San Sebastián: Springer, 2016: 211-222.

[17] MAULIK U, BANDYOPADHYAY S. Genetic algorithm-based clustering technique[J]. Pattern Recognition, 2000, 33(9): 1455-1465.

[18] BEZDEK J C, BOGGAVARAPU S, HALL L O, et al. Genetic algorithm guided clustering[C]//Proceedings of the first IEEE conference on evolutionary computation. IEEE world congress on computational intelligence. Orlando: IEEE, 1994: 34-39.

[19] JIMENEZ J F, CUEVAS F J, CARPIO J M. Genetic algo-

(下转第 172 页)