基于全加权矩阵分解的用户协同过滤推荐算法

邓秀勤[1] 刘太亨[1] 刘富春[2] 龙咏红[1]
(广东工业大学应用数学学院 广州 510006)[1] (广东工业大学计算机学院 广州 510006)[2]

摘 要 针对传统的基于用户协同过滤推荐算法将用户对某物品的喜好程度等同看待的问题,文中提出了一种融合全加权矩阵分解的用户协同过滤模型。该模型首先为观测值设计频率感知加权,且非均匀地设计用于未观测值的用户导向加权。然后组合观测值和未观测值的加权,并根据评分确定用户声誉和用户关系的相似性,构建融合全加权矩阵分解的用户协同过滤模型。为了验证提出的推荐算法的性能,在 Douban、Epinions 和 Last. fm 3 个真实数据集上进行了仿真实验。实验结果表明,所提出的 AWMF_UCFR 算法的推荐准确性与 MF 算法、WRMF-UO 算法、SoRS 算法相比有显著提高。

关键词 协同过滤,推荐算法,全加权矩阵分解,社交网络
中图法分类号 TP391 **文献标识码** A

User Collaborative Filtering Recommendation Algorithm Based on All Weighted Matrix Factorization

DENG Xiu-qin[1] LIU Tai-heng[1] LIU Fu-chun[2] LONG Yong-hong[1]
(School of Applied Mathematics, Guangdong University of Technology, Guangzhou 510006, China)[1]
(School of Computers, Guangdong University of Technology, Guangzhou 510006, China)[2]

Abstract Aiming at the problem that traditional user collaborative filtering recommendation algorithm equates users' preferences for an item, a user collaborative filtering model based on all weighted matrix decomposition was proposed. Firstly, the model designs frequency sensing weights for observations, and non-uniformly designs user-oriented weights for unobserved values. Then, the weights of the observed and unobserved values are combined, and the similarity between user reputation and user relationship is determined according to the score, and the user collaborative filtering model of the fused fully weighted matrix decomposition is constructed. In order to verify the performance of the proposed recommendation algorithm, experiments were carried out on three real data sets: Douban, Epinions and Last. fm. The experimental results demonstrate that the proposed AWMF_UCFR algorithm achieves significant improvements on recommendation accuracy than MF algorithm, WRMF-UO algorithm and SoRS algorithm.

Keywords Collaborative filtering algorithm, Recommendation algorithm, All-weighted matrix factorization, Social network

随着信息技术、大数据技术和人工智能技术的迅速发展,人们需要处理的数据量呈爆炸式增长。无论是在线访问还是移动应用程序访问,今天的用户面临着比以往任何时候都更加令人困惑的选择,推荐系统可以帮助用户在大型数据集中找到对他们最有价值的信息,是克服信息过载的一种有效途径。目前,应用范围最广且效果较好的推荐算法是协同过滤算法(Collaborative Filtering, CF),该方法是利用用户的相似性来推荐用户感兴趣的信息。由于 CF 不能有效地处理稀疏性和冷启动问题,导致协同过滤算法的推荐准确性不高。为了解决稀疏性和冷启动的问题,Agarwal 等[1]提出一种基于回归的隐语义模型,将响应预测为行和列潜在因子的乘法函数,通过对已知行和列特征的单独回归来估计并且提供一个统一的框架来解决冷启动和热启动问题。

Netflix Prize 比赛的诞生,催生出很多优秀的推荐算法,基于隐语义模型(Latent Factor Model, LFM)的推荐算法的研究取得了丰硕的成果,但传统的 LFM[2-3] 仅仅是单一地使用了用户对物品的评分信息,并没有考虑用户与用户之间的影响问题。对此,Qian 等[4]提出一种使用全局评分和局部评分相似性的社会推荐(SoRS),但是这种方法将不同用户的评分偏好程度都视为相同,忽略了不同用户对于同一物品偏好程度存在差异的问题。为了更好地解决隐语义模型在数据稀疏情况下推荐准确度低的问题,Deng 等[5]提出了一种新的隐语义模型推荐算法,该方法不仅考虑了用户对物品的评分信息和用户的社交关系信息,还考虑了物品信息,综合利用多种信息为推荐模型提供约束,而且采用潜在语义索引 LSI(Latent Semantic Index)技术来解决物品标签或类型数据的稀疏性问题,能更准确地找出潜在的相似物品,有效地提高了算法的推荐质量。Li 等[6]考虑到用户的访问频率不同,设计了观测数据的频率感知加权和非观测数据的用户定向加权,然后将观测数据和未观测数据的加权统一组合,得到用户评分全加权以替换 CF 算法进行快速优化,进而提出了全加权矩阵分解算法(WRMF-UO)。Lin 等[7]针对隐式反馈矩阵分解

本文受国家自然科学基金项目(61673122),广东省公益研究与能力建设专项资金资助项目(2015A030402006)资助。
邓秀勤(1966-),女,硕士,教授,主要研究方向为数据挖掘、智能计算,E-mail:dxq706@gdut. edu. cn(通信作者)。

(MF)推荐系统,提出了一种混合实时增量随机梯度下降(RISGD)更新技术,将权重正则化(ALSWR)[8]和随机梯度下降SGD[9]结合起来,并用具有隐式反馈的快速更新矩阵分解来解决推荐系统中隐式反馈数据的准确性问题。如今大多数现有的社交推荐系统都基于矩阵分解,而社交关系信息通常作为常规参数引入到基于模型的社交推荐系统中,譬如,Ma等[10]利用基于历史评分的用户之间的相似性作为推荐系统的局部影响,该框架可以增强用户之间的相互影响,提高推荐效果。

基于用户的协同过滤推荐一般是将不同用户的评分偏好程度都视为相同,却忽略了不同用户对于同一物品偏好程度可能存在差异的现实问题。为此,本文利用全加权方法对用户偏好程度进行优化,提出一种基于全加权矩阵分解的用户协同过滤推荐算法,首先将评分和用户社交引入到社交推荐系统,并设计加权正则化矩阵分解,为观测值设计频率感知加权,并非均匀地设计用于未观测值的用户导向加权。然后,以统一的方式组合观测值和未观测值的加权,构建全加权矩阵分解的用户协同过滤模型,以便更加准确地反映实际的推荐过程,从而有效地改进推荐算法的性能。在3个真实数据集上的实验结果表明,本文提出基于全加权矩阵分解的用户协同过滤推荐算法的性能优于3种对比算法。

1 相关研究工作

1.1 基于社交网络的推荐算法

社交网络的快速发展为解决协同过滤推荐算法中的数据稀疏和冷启动问题提供了良好的契机,研究人员充分利用社交网络提供的丰富信息来改进传统推荐算法的性能,提出了几种典型的基于社交网络的推荐算法,如SoRec,SocialMF,SoRS等。SoRec[11]是Ma等提出的基于概率矩阵分解模型的推荐算法,该算法集成了用户的评分信息和用户的社交网络信息,并通过用户评分信息和用户社交网络信息之间共享用户隐藏特征矩阵的方式来融合两种不同类型的信息源。SocialMF[12]是Jamali等提出的推荐算法,该算法在矩阵分解模型中集成了信任的传递机制来改进推荐算法的准确性。Qian等[4]提出了SoRS算法,将用户的社交关系,以及利用用户对物品的评分得到的用户推荐权威系数引入LFM中,同时修正了用户与好友的相似度计算公式,从而进一步提高了推荐准确度。实验结果表明SoRS算法的推荐准确度优于流行的SoRec算法和SocialMF算法。SoRS算法模型[4]如下:

$$F(U,V)=\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{m}I_{ij}C_i(r_{ij}-U_i^{\mathrm{T}}V_j)^2+\frac{\lambda_1}{2}\sum_{i=1}^{n}\sum_{k\in k(i)}sim(i,k)\|U_i-U_k\|_2^2+\lambda_2(\sum_{i=1}^{n}\|U_i\|_2^2+\sum_{j=1}^{m}\|V_j\|_2^2) \tag{1}$$

其中,C_i 表示用户对物品的偏好程度;r_{ij} 表示用户 i 对物品 j 的评分;$sim(i,k)$表示用户 i 和用户 j 的相似度;若用户 i 对物品 j 做出评分则 I_{ij} 为1,否则 I_{ij} 为0。

在现实生活中,不同用户对同一个物品的偏好程度有所差异,而基于用户的协同过滤算法却将不同用户的偏好程度看作一样的。这种假设可能不适用于现实生活场景,因为它们无法区分整个数据的贡献,并且很容易导致预测偏差。因此本文将运用全加权矩阵方法对用户偏好差异进行优化。

1.2 全加权矩阵分解算法

矩阵分解被广泛运用到CF推荐,特别是显式反馈和隐式反馈的CF推荐。现有的大多数CF推荐的策略有两种,一种是逐点回归,另一种是成对排序策略。然而大多数现有逐点回归研究的一个主要局限是对未观测值均匀加权,这限制了模型的有效性和可扩展性。成对排序策略试图将观测值置于未观测值之上,忽略了表示偏好置信度的绝对有用的数值,而且还以不切实际的方式对未观测值进行建模。针对这个问题,Li等[6]提出全加权矩阵分解算法(WRMF-UO),将用户访问频率的差异考虑到协同过滤算法中,其模型框架如下:

$$L(U,V)=\sum_{u=1}^{M}\sum_{i\in R_u}C_{ui}(1-\hat{R}_{ui})^2+\sum_{u=1}^{M}\sum_{i\in R_u}d_u(0-\hat{R}_{ui})^2+\lambda(\sum_{u=1}^{M}\|U_u\|_2^2+\sum_{i=1}^{N}\|V_i\|_2^2) \tag{2}$$

其中,C_{ui} 表示观测值的权重;d_u 表示未观测值的权重;$\hat{R}_{ui}$ 是矩阵分解后的预测得分。

WRMF-UO算法虽然考虑到用户对物品的偏好程度是不均匀的,但是忽略了用户之间的关系。因为一个用户能够直接或者间接地影响到另一个用户对同一物品的喜好程度,所以本文提出融合全加权矩阵分解与用户协同过滤的推荐算法来提高推荐的准确性。

2 融合全加权矩阵分解与用户协同过滤的推荐算法

2.1 算法模型

本节将给出融合全加权矩阵分解与用户协同过滤的推荐算法模型。首先将用户 u 消费过物品 $i(r_{ui}>0)$,即用户 u 偏爱物品 i 称为观测值[8];否则,将用户 u 不喜欢物品 i 称为未观测值。

在传统的推荐方法上,很多学者都忽略了访问频率的差异,并通过给出均匀的权重来均等地处理观测值[2-3,13-14]。这种处理方式忽略了重要的偏好信息,影响了推荐的准确性。为了提高推荐的准确性,本文将用户的访问频率作为观测值并为其分配了权重,具体如下:

$$D_i=\alpha|r(i)|\frac{g_i^{\tau}}{\sum_{i\in r(i)}g_i^{\tau}} \tag{3}$$

其中,α 确定观测值的全局权重。g_i 表示用户访问物品 i 的次数。为了增强推荐的灵活性和有效性,添加了 τ 控制频繁物品的置信水平。当 $\tau=0$,D_i 为均匀的权重;当 $\tau\neq0$,频繁物品的置信加强,以区分非频繁物品。

对于隐式反馈,其中很少的一部分是观测值,大部分是未观测值。对于未观测值,本文提出一种非均匀的面向用户的加权,其假定用户喜欢一种物品的概率很大时,用户更可能不喜欢其他的物品,即该用户的未观测值的概率更加否定。未观测值的权重定义如下:

$$E_i=\beta\frac{|\lfloor r(i)\rfloor|^{\pi}}{\sum_{i=1}^{M}|\lfloor r(i)\rfloor|^{\pi}} \tag{4}$$

其中,β 是控制未观察数据总重量的系数;$\lfloor r(i)\rfloor$ 是用户观察到的项目数。同样添加 π 来控制活跃用户的级别。当 $\pi=0$ 时,$E_i=\frac{\beta}{M}$ 表示对未观察到的数据进行统计加权;当 $\pi\neq0$ 时,不均匀加权 E_i 由 π 控制。

其次,对于推荐系统来说,用户社交关系会影响推荐系统的推荐准确性。一个用户能够直接或者间接地影响到另一个

用户对同一物品的喜好程度。用户社交关系相似度[15]可以表示为:

$$sim(i,k)=\frac{\sum_{z\in N(i)\cap N(k)}\frac{1}{\log(1+|N(z)|)}}{\sqrt{|N(i)||N(k)|}} \tag{5}$$

其中,$N(i)$表示用户 i 购买过物品的集合;$N(k)$表示用户 k 购买同一物品的集合;$N(z)$表示用户 i 和用户 k 共同购买同一物品的集合。

最后,为了有效地利用观测值和未观测值以及用户的社交关系,本文提出一种新的基于全加权矩阵分解的用户协同过滤模型(User Collaborative Filtering Recommendation Based on All-Weighted Matrix Factorization, AWMF_UCFR),该模型旨在解决用户对物品的偏爱不均匀的优化问题,该模型表示如下:

$$F(U,V)=\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{m}I_{ij}D_i(1-U_i^{\mathrm{T}}V_j)^2+\frac{1}{2}\sum_{\lfloor r(i)\rfloor=1}^{M}\sum_{i\in I/U_i}E_i(0-U_i^{\mathrm{T}}V_j)^2+\frac{\lambda_1}{2}\sum_{i=1}^{n}\sum_{k\in k(i)}sim(i,k)\|U_i-U_k\|_2^2+\lambda_2\left(\sum_{i=1}^{n}\|U_i\|_2^2+\sum_{j=1}^{m}\|V_j\|_2^2\right) \tag{6}$$

其中,用户 i 对 j 做出评分则 I_{ij} 为 1,否则 I_{ij} 为 0。应用梯度下降法对目标函数(6)进行求解,可得梯度为式(7)和式(8)。

$$\frac{\partial F}{\partial U_i}=\sum_{j=1}^{m}I_{ij}D_i(1-U_i^{\mathrm{T}}V_j)(-V_j)+\sum_{i\in I/U_i}E_i(0-U_i^{\mathrm{T}}V_j)(-V_j)+\lambda_1\sum_{k\in k(i)}sim(i,k)(U_i-U_k)+\lambda_2U_i \tag{7}$$

$$\frac{\partial F}{\partial V_j}=\sum_{i=1}^{n}I_{ij}D_i(1-U_i^{\mathrm{T}}V_j)(-U_j)+\sum_{\lfloor r(i)\rfloor}^{M}E_i(0-U_i^{\mathrm{T}}V_j)(-U_j)+\lambda_2V_j \tag{8}$$

然后可以通过式(9)和式(10)计算得到更新后的 U 和 V。

$$U_i=U_i-\Delta\left[\sum_{j=1}^{m}I_{ij}D_i(1-U_i^{\mathrm{T}}V_j)(-V_j)+\sum_{i\in I/U_i}E_i(0-U_i^{\mathrm{T}}V_j)(-V_j)+\lambda_1\sum_{k\in k(i)}sim(i,k)(U_i-U_k)+\lambda_2U_i\right] \tag{9}$$

$$V_j=V_j-\Delta\left[\sum_{i=1}^{n}I_{ij}D_i(1-U_i^{\mathrm{T}}V_j)(-U_j)+\sum_{\lfloor r(i)\rfloor}^{M}E_i(0-U_i^{\mathrm{T}}V_j)(-U_j)+\lambda_2V_j\right] \tag{10}$$

其中,Δ 表示下降速度。

2.2 算法描述

算法 1

输入:用户评分矩阵 R,用户关系矩阵 K,潜在因子矩阵的维数 L,权重矩阵 w,正则化参数 λ_2,最大迭代次数 n

输出:预测评分矩阵 $\hat{R}$

1. 初始化:初始化用户矩阵 U~N(0,0.1)与物品矩阵 V~N(0,0.1)。

2. 计算预测评分矩阵 $\hat{R}=U_iV_j^{\mathrm{T}}(\forall(i,j)\in R_0)$

```
while 不收敛或迭代小于最大迭代次数 n do
  for u=1 to M
    for k=1 to K
      通过式(9)更新 Ui;
      更新 R^
    endfor
  endfor
  预计算用户的相似度 Q=U^T d̃U;
  for u=1 to N
    for k=1 to K
      通过式(10)更新 Vj;
      更新 R^;
    endfor
  endfor
end
```

3. 输出预测评分矩阵 $\hat{R}$。

2.3 复杂度分析

观测数据的加权由式(3)交互计算,式(3)、式(4)的时间复杂度是 O(n)和 O(n)。从式(5)可以看出,$sim(i,k)$的计算复杂度为 O($m\cdot n$)。实际上,$sim(i,k)$的值通常是预先离线计算的。梯度法的主要计算是目标函数的 F 及其对变量 U 和 V 的梯度。由于矩阵 R 的稀疏性,梯度 $\frac{\partial F}{\partial U_i}$ 和 $\frac{\partial F}{\partial V_j}$ 的计算复杂度都是 O(θk),其中 θ 是矩阵 R 中的非零项的数量,k 是用户/物品潜在因子向量的维度。评估目标函数 $F(U,V)$的计算复杂度为 O($n+\theta k$)。

3 实验与结果分析

为了验证本文提出的 AWMF_UCFR 算法的性能,本文在 3 个真实的数据集上将所提算法与 MF 算法[7]、SoRS 算法[4]和 WRMF-UO 算法[6]进行实验并对算法性能做对比分析。

3.1 数据集

实验中采用 Douban[10]、Epinions[14]和 Last.fm[16] 3 个数据集。这些数据集在相关参考文献的主页公开获得。本文从 Douban 数据集中选取 7 700 名用户,并将这些用户作为主要研究对象,进一步获得用户的社交网络和电影评分。在 Epinions 数据集中,本文选取评分次数至少为 3 次的 22 168 名用户作为研究对象。Last.fm 数据集是来自 Last.fm 在线音乐平台用户-艺术家收听信息。本文从 1 892 个用户和 17 632 个艺术家中获得 92 834 个评分结果。表 1 列出了这 3 个数据集的统计情况。

表 1　数据集 Douban,Epinions 和 Last.fm 的统计数据

数据集	Douban	Epinions	Last.fm
用户数	7 700	22 168	1 892
物品数	6 497	41 369	17 632
物品评分数	851 220	583 968	92 834
评分稀疏度	0.017	0.000 636	0.002 8
好友关系	5 874	342 037	12 717
好友关系稀疏度	0.000 198 17	0.000 696	0.003 6
聚类系数	0.063	0.176	0
共同评分关系	4 119	16 196	11 946

3.2 评价指标

为了对模型进行评价,本文采用平均绝对误差(MAE)作为衡量推荐算法预测准确度的指标。MAE 定义如下:

$$MAE=\frac{1}{|F^{Test}|}\sum_{(i,j)\in F^{Test}}(r_{ij}-\hat{r}_{ij})^2 \tag{11}$$

其中,r_{ij} 表示用户 i 对物品 j 的评分,$\hat{r}_{ij}$ 表示户 i 通过算法对物品 j 的预测评分。F^{Test} 表示具有实际评分的(用户,项目)对的集合;$|F^{\mathrm{Test}}|$ 表示测试集中的实际评分数。

3.3 模型参数的影响

从目标函数(6)可以看出,本文需要确定 α,τ,β,π 和 λ_1 5 个参数。参数 α,τ 与观察值的权重相关联:参数 α 用来控制观测值的全局权重,τ 表示频繁项目超过非频繁项目的比重。

参数 β,π 考虑了未观测值的权重:β 反映了未观测值的总体权重,π 控制活动用户的评分而不是其他非活动用户的评分。参数 λ_1 是控制相似用户对推荐结果影响的比重。不同参数对 Douban 数据集、Epinions 数据集和 Last.fm 数据集的影响分别如图 1—图 15 所示。

从图 1 至图 5 可以看出,在 Douban 数据集中,当 $\alpha=4$,$\tau=0.7$,$\beta=1500$,$\pi=1$,$\lambda_1=0.001$ 时,MAE 值最小。

从图 6 至图 10 可以看出,在 Epinions 数据集中,当 $\alpha=5$,$\tau=0.6$,$\beta=1000$,$\pi=0.6$,$\lambda_1=0.0001$ 时,MAE 值最小。

从图 11 至图 15 可以看出,在 Last.fm 数据集中,当 $\alpha=5$,$\tau=0.6$,$\beta=1000$,$\pi=0.6$,$\lambda_1=0.0001$ 时,MAE 值最小。

在数据集 Douban 上,先将参数 α 和 τ 的权重固定[17]为 1,如图 1、图 2 所示。这意味着平等地处理未观察数据。然后研究面向用户的加权策略如何影响推荐性能。从图 1 可以看出,通过设置 $\tau=0$(这意味着统一加权),AWMF_UCFR 的性能随 α 的变化而变化。当 $\alpha=4$ 时,MAE 值最小,这个结果意味着 α 的重要性,它决定了观测数据的全局权重,也就是说 $\alpha=4$ 时 AWMF_UCFR 算法推荐的性能最好。然后设定 $\alpha=4$,研究 τ 如何影响推荐结果。从图 2 可以看出,随着 τ($\tau\in[0,1]$)的增加,AWMF_UCFR 的性能逐渐提高,之后 τ 下降到[0.7,0.8]附近。τ 的最佳值位于[0.7,0.8]附近,也就是推荐性能优于均匀权重,这揭示出传统推荐算法的加权策略的缺点,也进一步验证 AWMF_UCFR 算法面向用户的观测数据加权的有效性。

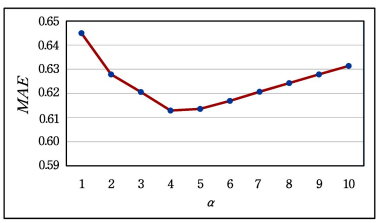

图 1 在 Douban 数据集中参数 α 的影响

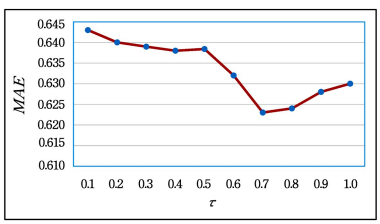

图 2 在 Douban 数据集中参数 τ 的影响

接下来,将观测数据的 α 和 τ 的权重分别设定为 4 和 0.7,然后探讨对全数据权重的充分考虑是否比只考虑观察到的权重的情况更好。首先我们设定 $\pi=0$,这表示未观测到的数据的统一加权,然后观察 β 的变化,如图 3、图 4 所示。可以看出,MAE 的值随着 β 的变化而变化,并在 $\beta=1500$ 处 MAE 的值最小。该结果反映了未观察到的数据的总体权重的重要性。然而,与图 2 相比,只获得了一点点增益。这是因为它们都在未观察到的数据上分配了均匀的权重,图 3 只是找到了与图 2 相当的非观测数据的次优整体权重。设定 $\beta=1500$,再研究参数 π 如何影响 AWMF_UCFR 的推荐性能。从图 4 可以看出随着 $\pi\in[0,1]$的增加,MAE 的值逐渐降低,也就是 AWMF_UCFR 的推荐性能增加,并且当 $\pi=1$ 时,推荐效果最佳,此结果反映了将活动用户正项加权项目作为否定项的重要性。最后,分析 λ_1 如何影响 AWMF_UCFR 的推荐精度,从图 5 可以看出,随着 λ_1 的下降,MAE 的值由大变小再由小变大。当 $\lambda_1=0.001$ 时,MAE 的值最小,也就是 AWMF_UCFR 的推荐精度最高。

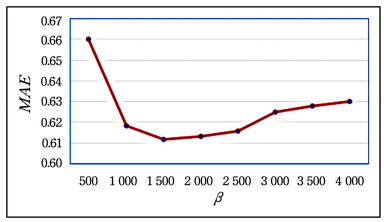

图 3 在 Douban 数据集中参数 β 的影响

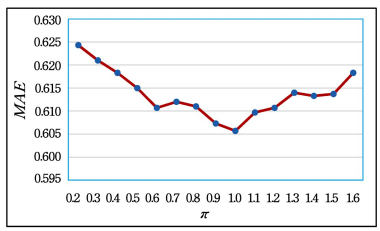

图 4 在 Douban 数据集中参数 π 的影响

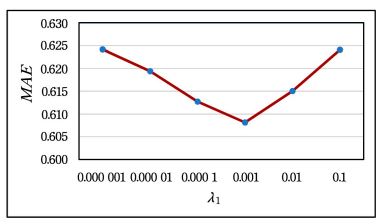

图 5 在 Douban 数据集中参数 λ_1 的影响

在 Epinions 数据集上(如图 6 至图 10 所示)和在 Last.fm 数据集上(如图 11 至图 15 所示)全数据加权的影响与 Douban 数据集的结果相似,在此不再详述。

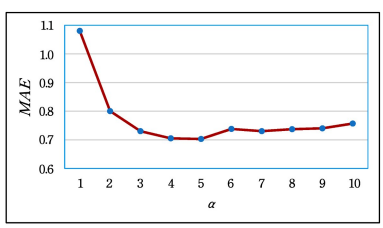

图 6 在 Epinions 数据集中参数 α 的影响

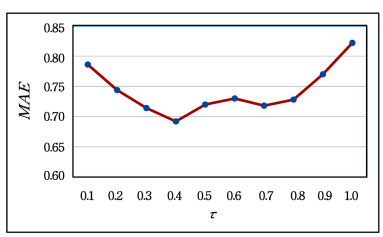

图 7 在 Epinions 数据集中参数 τ 的影响

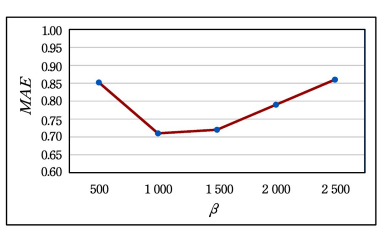

图 8 在 Epinions 数据集中参数 τ 的影响

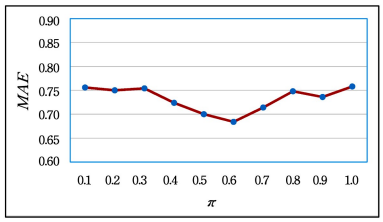

图 9　在 Epinions 数据集中参数 π 的影响

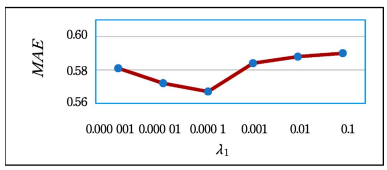

图 10　在 Epinions 数据集中参数 λ_1 的影响

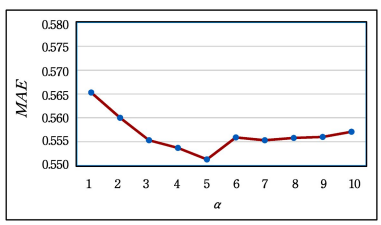

图 11　在 Last. fm 数据集中参数 α 的影响

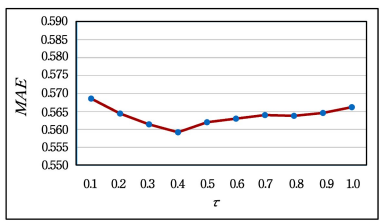

图 12　在 Last. fm 数据集中参数 τ 的影响

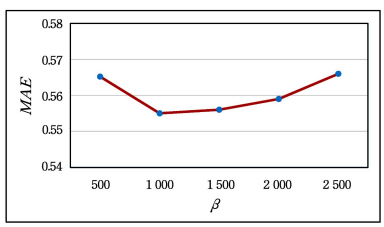

图 13　在 Last. fm 数据集中参数 β 的影响

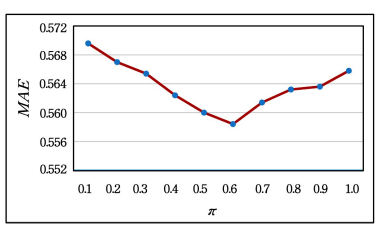

图 14　在 Last. fm 数据集中参数 π 的影响

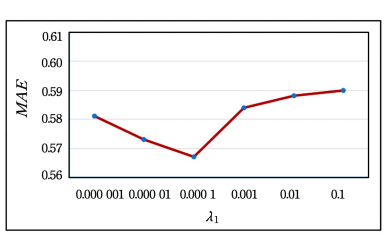

图 15　在 Last. fm 数据集中参数 λ_1 的影响

3.4　实验结果分析

在实验过程中，选取 Douban[10]，Epinions[14] 和 Last. fm[16] 3 个数据集。将本文提出的算法（AWMF_UCFR）和 MF[7]，SoRS[4]，WRMF-UO 算法[6] 在 3 个数据集上进行实验，并比较各个算法的推荐性能和计算复杂度。

1）推荐性能分析

表 2 显示了本文所提出的算法 AWMF_UCFR 和 MF 算法、SoRS 算法、WRMF-UO[6] 算法推荐性能的对比，其中在 SoRS 算法中参数 λ_1 设为 0.001。可以看出，本文提出的 AWMF_UCFR 算法在 Douban 和 Last. fm 这两个数据集上的 *MAE* 都比 MF 算法、SoRS 算法和 WRMF-UO 算法要小，而且推荐精度分别提高了 1.09%，6.71%，3.73%。

表 2　4 种方法在 Douban，Epinions，Last. fm 数据集上的性能比较

数据集	评价指标	MF	SoRS	WRMF-UO	AWMF_UCFR
Douban	*MAE*	0.5746	0.5647	0.5324	**0.5218**
	Improve/%	1.72	5.72	2.0	
Epinions	*MAE*	0.8812	0.8753	0.7836	**0.7342**
	Improve/%	0.67	10.47	6.30	
Last. fm	*MAE*	0.6148	0.6094	0.5853	**0.5689**
	Improve/%	0.88	3.95	2.80	

2）计算复杂度分析

由 3.3 节可知，本文 AWMF_UCFR 算法的计算复杂度为 $O(n+\theta k)$；在文献[4]中 SoRS 算法的计算复杂度为 $O(n+\theta k)$；在文献[6]中 WRMF-UO 算法的计算复杂度为 $O(mnk^2)$；在文献[7]中 MF 算法的计算复杂度为 $O(\theta k^2+(n+m)k^2)$。由此可见，本文提出的 AWMF_UCFR 算法在计算复杂度上等于 SoRS 算法，但优于 MF 算法和 WRMF-UO 算法。

结束语　本文提出一种融合全加权矩阵分解与用户协同过滤的推荐算法。与之前的 MF 算法、SoRS 算法和 WRMF-UO 算法相比，无论是对未观测值分配了统一权重还是对观测值应用了固定权重，全加权方案都充分考虑了整个数据，并为观测值和未观测值分配了可用的置信权重。在 3 个真实数据集上的实验结果表明，本文提出的方法都优于其他的算法。

对于未来的工作，我们将主要在两个方面扩展本文的方法：1）研究如何利用辅助数据（例如物品内容）来进一步改进加权并提高推荐的准确性；2）研究如何将基于回归的方法与基于排序的方法（即贝叶斯个性化排名）相结合，充分利用异构损失的优势来获得更好的推荐性能。

参 考 文 献

[1] AGAWAL D，CHEN B C. Regression-based Latent Factor Models[C]// 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Paris，France：ACM，2009：19-27.

[2] 王升升，赵海燕，陈庆奎，等. 个性化推荐中的隐语义模型[J]. 小型微型计算机统，2016，37(5)：881-889.

[3] 范慧婷，钟春琳，龚海华. 基于隐语义模型的个性化推荐[J]. 计算机应用与软件，2017，34(12)：206-210.

[4] QIAN F，ZHAO S，TANG J，et al. SoRS：Social recommendation using global rating reputation and local rating similarity [J]. Physica A Statistical Mechanics & Its Applications，2016，461：61-72.

（下转第 219 页）

较高准确度，充分验证了 1NN-kmeans 算法理论上的可行性和有效性。由图 4 可以看出，1NN-kmeans 运行时间稍长，这是因为计算 $r_i<\varepsilon$ 的数据样本的最近邻耗费了时间，但 3 种算法的运行时间都在毫秒级，均未超过 1 s。

1NN-kmeans 算法的聚集性较好，聚类的结果通常只是某一簇的某些数据样本错误地全部分到另一个簇中，而其他的簇能够全部分类正确，通常不会是原本一个簇中的数据被分到其他几个簇中去。1NN-kmeans 聚类结果很稳定，在数次实验中能够得到不变的结果，对某些数据集的准确度能够达到 100%。针对数据样本与簇的关系不同采用不同的划分方法更合理和有效。

结束语　传统的 k-means 算法将所有属于簇的程度不同的数据样本同等对待，一概按照最小距离原则划分。本文提出的 1NN-kmeans 算法通过计算第二最小距离与第一最小距离的比值是否大于某一参数 ε，来界定数据样本是否“很属于”某一簇，而与其他簇关系很小，对于不是“很属于”簇的数据样本，采用近邻思想将其划分到最近邻所在的簇中。实验选择 UCI 数据库中的 6 个数据集，对算法的有效性进行验证。通过与 k-means，k-means＋＋对比可以得出，1NN-kmeans 算法能够在较低的迭代次数中，达到较高的准确度和较低的误差平方和，聚类结果稳定，适应不同簇规则的数据集，是一种高效的 k-means 优化算法。

参 考 文 献

[1] 高曼，韩勇，陈戈，等. 基于 K-means 聚类算法的公交行程速度计算模型[J]. 计算机科学，2016，43(S1)：422-424，439.

[2] 赵建民，管国权，王红艳. 基于遗传算法的硬聚类算法改进[J]. 计算机工程与科学，2008(8)：83-85.

[3] 唐胡鑫. 电子商务客户忠诚度模型仿真研究[J]. 计算机仿真，2016，33(1)：413-415，424.

[4] 王勇，唐靖，饶勤菲，等. 高效率的 K-means 最佳聚类数确定算法[J]. 计算机应用，2014，34(5)：1331-1335.

[5] 谢娟英，王艳娥. 最小方差优化初始聚类中心的 K-means 算法[J]. 计算机工程，2014，40(8)：205-211，223.

[6] 郁启麟. K-means 算法初始聚类中心选择的优化[J]. 计算机系统应用，2017，26(5)：170-174.

[7] 邢长征，谷浩. 基于平均密度优化初始聚类中心的 k-means 算法[J]. 计算机工程与应用，2014，50(20)：135-138.

[8] 朴尚哲，超木日力格，于剑. 模糊 C 均值算法的聚类有效性评价[J]. 模式识别与人工智能，2015，28(5)：452-461.

[9] 马闯，吴涛，段梦雅. 基于 K 近邻隶属度的聚类算法研究[J]. 计算机工程与应用，2016，52(10)：55-58，117.

[10] 王超学，潘正茂，马春森，等. 改进型加权 KNN 算法的不平衡数据集分类[J]. 计算机工程，2012，38(20)：160-163，168.

[11] 华辉有，陈启买，刘海，等. 一种融合 Kmeans 和 KNN 的网络入侵检测算法[J]. 计算机科学，2016，43(3)：158-162.

[12] 苏毅娟，邓振云，程德波，等. 大数据下的快速 KNN 分类算法[J]. 计算机应用研究，2016，33(4)：1003-1006，1023.

[13] ARTHUR D，VASSILVITSKII S. k-means＋＋：the advantages of careful seeding[C]// Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms. Society for Industrial and Applied Mathematics Philadelphia，PA，USA，2007：1027-1035.

[14] 余秀雅，刘东平，杨军. 基于 K-means＋＋的无线传感网分簇算法研究[J]. 计算机应用研究，2017，34(1)：181-185.

[15] ASUNCION A，NEWMAN D J. UCI machine learning repository[EB/OL]. [2009-12-23]. http://archive.ics.uci.edu/.

（上接第 203 页）

[5] DENG X Q，LIU T H，LI W Z，et al. A Latent Factor Model of Fusing Social Rregularization Term and item Regularization Term [J]. Physic A：Statistical Mechanics and its Applications，2019，525：1330-1342.

[6] LI H，DIAO X，CAO J，et al. Collaborative Filtering Recommendation Based on All-Weighted Matrix Factorization and Fast Optimization [J]. IEEE Access，2018，6：25248-25260.

[7] LIN C，WANG L，TSAI K. Hybrid Real-Time Matrix Factorization for Implicit Feedback Recommendation Systems [J]. IEEE Access，2018，6(10)：21369-21380.

[8] HU Y，KOREN Y，VOLINSKY C. Collaborative filtering for implicit feedback datasets[C]// Proceedings of the 8th IEEE International Conference on Data Mining. Pisa，Italy：IEEE，2008：263-272.

[9] LING G，YANG H，KING I，et al. Online learning for collaborative filtering[C]// The 2012 International Joint Conference on Neural Networks. Brisbane，Australia：ACM，2012：1-8.

[10] MA H，ZHOU D，LIU C，et al. Recommender systems with social regularization[C]// Proceedings of the Fourth ACM International Conference on Web Search and Data Mining. Hong Kong：ACM，2011：287-296.

[11] MA H，YANG H，LYU M R，et al. Sorec：social recommendation using probabilistic matrix factorization [C]// Proceedings of the 17th ACM Conference on Information and Knowledge Management. Napa Velley，Califomia：ACM，2008：931-940.

[12] JAMALI M，ESTER M. A matrix factorization technique with trust propagation for recommendation in social networks[C]// Proceedings of the 4th ACM Conference on Recommendation Systems. Barcelona，Spain：ACM，2010：135-142.

[13] PAN R，ZHOU Y H，CAO B. One-class collaborative filtering [C]// Proceedings of the 8th IEEE International Conference on Data Mining. Pisa，Italy：IEEE，2008：502-511.

[14] TANG J，HU X，GAO H，et al. Exploiting local and global social context for recommendation[C]// Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence. Palo Alto，Califomia：AAAI，2013：2712-2718.

[15] BREESE S J，HECKERMAN D，KADIE C M. Empirical Analysis of Predictive Algorithms for Collaborative Filtering[C]// Proceedings of the Fourteenth Annual Conference on Uncertainty in Artificial Intelligence. Madison，wisconsin：ACM，1998：43-52.

[16] MINAS G，CARTER T B，MACIEJ K，et al. Multigraph Sampling of Online Social Networks [J]. IEEE Journal on Selected Areas in Communications，2011，29(3)：1893-1905.

[17] LIAN D，ZHAO C，XIE X，et al. GeoMF：Joint geographical modeling and matrix factorization for point-of-interest recommendation[C]// Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York：ACM，2014：831-840.