Pure Mathematics 理论数学, 2021, 11(5), 929-936
Published Online May 2021 in Hans. http://www.hanspub.org/journal/pm
https://doi.org/10.12677/pm.2021.115106

Hans汉斯

基于改进聚类与融合用户属性特征的协同过滤推荐算法

王汇琳，陈　欣*

沈阳工业大学理学院，辽宁 沈阳
Email: *849923895@qq.com

收稿日期：2021年4月17日；录用日期：2021年5月20日；发布日期：2021年5月27日

摘　要

针对协同过滤算法数据稀疏导致推荐质量低和推荐效率低的问题，本文提出了一种基于改进K-means聚类与用户属性的协同过滤推荐算法。为了改进K-means算法初始中心选取的随机性，本文先用canopy算法对数据进行粗聚类，引入“最大最小距离积法”选取初始点，接着用K-means算法进行聚类，在生成多个聚类簇之后，将修正的余弦相似度与用户属性特征相结合，形成新的相似度计算模型，最后进行相应的推荐。通过MAE、RMSE两个指标的比较，结果表明，改进后的算法能够提高推荐效率和推荐准确性。

关键词

K-Means聚类算法，协同过滤，最大最小距离积法，最近邻用户

Collaborative Filtering Recommendation Algorithm Based on Improved Clustering and Fusion of User Attribute Features

Huilin Wang, Xin Chen*

School of Science, Shenyang University of Technology, Shenyang Liaoning
Email: *849923895@qq.com

Received: Apr. 17th, 2021; accepted: May 20th, 2021; published: May 27th, 2021

*通讯作者。

文章引用: 王汇琳, 陈欣. 基于改进聚类与融合用户属性特征的协同过滤推荐算法[J]. 理论数学, 2021, 11(5): 929-936.
DOI: 10.12677/pm.2021.115106

Abstract

In order to solve the problem of low recommendation quality and low recommendation efficiency, which is caused by data sparseness in collaborative filtering algorithm, a collaborative filtering recommendation algorithm based on improved K-means clustering and user attribute was proposed in order to improve the randomness of initial center selection of K-means algorithm. In this paper, Canopy algorithm was used to perform crude clustering of data, and "maximum and minimum distance product method" was introduced to select initial points. Then, K-means algorithm was used for clustering. After the generation of multiple clustering clusters, the revised cosine similarity and user attribute characteristics are combined to form a new similarity calculation model. Finally, the corresponding recommendation is made. Through the comparison of MAE and RMSE, the results show that the improved algorithm can improve the efficiency and accuracy of recommendation.

Keywords

K-Means Clustering Algorithm, Collaborative Filtering, Maximum and Minimum Distance Product Method, Nearest Neighbor User

1. 引言

随着网络的快速发展，用户和商品数量不断增加，信息过载的问题越加显著，因此，推荐系统应运而生。协同过滤算法[1] [2]是推荐系统中最常用的推荐算法。此外，传统的协同过滤存在数据稀疏、冷启动、推荐结果准确率低等问题。针对上述不足，许多学者都对协同过滤算法进行了改进。唐泽坤[3]等在传统 canopy 聚类算法的基础上做了改进，较好地解决了用户聚类的问题。郑杰[4]等对人工蜂群算法进行改进，再将其与 K 均值算法有效地结合，最后证明了改进算法的有效性。向晓东[5]等通过项目相似度筛选出待预测评分项目的近邻用户集，在计算项目间偏差值时引入用户相似度，从而有效地提高了评分预测的可靠性。本文依据上述学者的思路，提出一种基于改进 K-means 聚类与融合用户属性特征的协同过滤推荐算法，该算法通过建立混合聚类模型进行聚类，接着将修正余弦相似度与用户属性特征进行融合形成一种新的相似度计算算法，最后给出推荐，进行实证分析，验证该算法的推荐效果。

2. 相关工作

2.1. Canopy 聚类算法

Canopy 是一种“粗”聚类算法[6]，相比其他聚类算法，它的优点在于得到簇的速度很快，抗噪能力强，简单易实现，但是精度较低，分类结果不稳定，可以结合 K-means 算法一起使用。

Canopy 聚类算法步骤如下：

输入：n 个对象的数据集 D

1) 确定两个初始距离阈值 t_1、t_2 ($t_1 > t_2$)。

2) 从数据集 D 中随机选取一个数据 Q，计算这个数据 Q 到所有 canopy 的距离。

3) 如果当前没有 canopy，则直接将 Q 作为 canopy 中心点，并将其从 D 中删除。

4) 如果 Q 与某个 canopy 的距离小于 t_1，则将 Q 标上弱标记并添加到这个 Canopy 中，同时需要把 Q 从 D 中删除(这个数据可以作为新的 canopy 来计算其他数据到这个点的距离)。

5) 如果与某个 Canopy 距离小于 t_1，大于 t_2，同样将 Q 加入到这个 Canopy，但不将其从 D 中删除。

6) 如果 Q 与某个 canopy 的距离小于 t_2，则将 Q 标上强标记并添加到这个 Canopy 中，同时需要把 Q 从 D 中删除。此时认为这个数据点距离该 canopy 已经足够近了，不需要形成新的 canopy。

7) 循环步骤 2)~6)，直至数据集 D 中没有数据。

输出：k 个聚类中心。

2.2. K-Means 聚类算法

K-means 聚类算法[7]简单易实现，聚类效果好具体算法流程如下：

输入：原始数据样本集合 X

1) 随机选取 k 个点作为为初始聚类中心。

2) 分别计算每个对象到聚类中心的距离，根据最小距离原则分配到最近的聚类簇中，并更新聚类中心。

3) 重新计算 k 个新聚类中心。

4) 若当前中心点和原中心点之间的距离小于设置的阈值，即可认为聚类已达到所期望的结果，算法终止。

5) 否则需要重复步骤 2)~步骤 4)直到聚类中心不再发生变化。

输出：k 个簇的集合。

2.3. 修正余弦相似度

$$sim_{a\cos}(u,v)=\frac{\sum_{i\in I_{uv}}\left(r_{u,i}-\overline{r}_u\right)\left(r_{v,i}-\overline{r}_v\right)}{\sqrt{\sum_{i\in I_u}\left(r_{u,i}-\overline{r}_u\right)^2}\sqrt{\sum_{i\in I_v}\left(r_{v,i}-\overline{r}_v\right)^2}} \tag{1}$$

其中 $I_{u,v}$ 是用户 u 和用户 v 的共同评分项目集，$r_{u,i}$ 和 $r_{v,i}$ 分别表示用户 u 和用户 v 对项目 i 的评分，$\overline{r}_u$ 是用户 u 的平均评分，$\overline{r}_v$ 是用户 v 的平均评分。

3. 基于改进聚类的协同过滤算法

3.1. 改进 K-Means 聚类算法描述

先由 Canopy 算法进行粗聚类得到 k 个聚类中心，为了改进 K-means 算法初始点选取的随机性，本文引入“最大最小距离积法”对其进行优化，用来提升准确度，它的优点是算法所需参数少，使用最大最小距离乘积能选取到密度较大的点，稀疏初始点分布，并且大概率地可以避免出现区域内点密度相差很大而两个距离积相等的情况，同时点与点之间的差异可以用乘积来放大，使得选取的过程更具区分度。克服原始算法初始化的随机性。避免初始中心的选取过于稠密，从而出现聚类冲突的现象，并利用更新公式进行迭代寻优。再用 K-means 算法进行聚类。

3.2. 基于改进 K-Means 聚类的流程

改进 K-means 聚类算法步骤如下：

输入：n 个对象的数据集 D

1) 用 Canopy 算法进行聚类得到 k 值。

2) 从集合 D 中随机选取点作 Z_1 为第一个初始点，将其加入集合 Z 中，并从集合 D 中删除。

3) 计算更新后 D 中所有元素到 Z_1 的距离，选取距离 Z_1 最大的点为 Z_2。

4) 将此点加入集合 Z 中，并从 D 中删除。

5) 分别计算更新后 D 中元素到 Z 中各个元素的距离并存入 Temp 中。

6) 计算 D 中每个元素对应的 Temp 的最大值与最小值的乘积(max (Texp) × min (Temp))，取该值最大对应点。

7) 如果选取的初始点个数 < k，则循环 2)~6)，直至输出包含 k 个初始点的集合 Z。

8) 进行 K-means 聚类，得到最终聚类的结果。

输出：多个聚类簇

其中：D 是包含所有数据的集合；k 是要选取的初始点个数；Z 是存储待加入的 k 个初始点的集合，算法开始前为空集；Temp 是存储 Z 中各个元素到 D 中各个元素乘积结果的数组。

3.3. 融合用户特征相似度算法

在进行推荐时，由于性别，年龄，职业等不同，用户在选择项目时往往也不同，相同用户属性的人喜欢的项目也具有一定的相似性。传统的相似度计算方法经常会忽略不同用户属性的相似性，因此，本文将用户性别与年龄特征融入修正余弦相似度中。

1) 年龄属性相似度

用户 u 与用户 v 之间的年龄相似度如下：

$$N(u,v)=e^{-|n_u-n_v|} \tag{2}$$

式中：$N(u,v)$ 取值范围为 $[0,1]$ 之间，值越大，相似度越大；n_u 为用户 u 的年龄；n_v 为用户 v 的年龄。

2) 性别属性相似度。

用户 u 与用户 v 之间的性别相似度如下：

$$X(u,v)=\begin{cases}0, & X_u \neq X_v \\ 1, & X_u = X_v\end{cases} \tag{3}$$

式中：X_u 为用户 u 的性别；X_v 为用户 v 的性别。

3) 用户属性相似度。综合上述年龄属性相似度、性别属性相似度，得出用户属性相似度如下：

$$sim_{NX}(u,v)=\alpha N(u,v)+(1-\alpha)X(u,v) \tag{4}$$

式中 $\alpha \in [0,1]$ 为属性的权重系数，在不同的推荐系统中，可以采用回归分析拟合对其调整。将用户属性相似度与修正余弦相似度相结合，可以得到一种新的相似度计算模型，即融合用户特征相似度计模型，如下：

$$sim_a(u,v)=\beta sim_{NX}(u,v)+(1-\beta)sim_{a\cos}(u,v) \tag{5}$$

式中 $\beta \in [1,0]$ 为权重系数。

3.4. 基于改进 K-Means 聚类和融合用户属性的协同过滤推荐算法

由于 K-means 算法聚类数和初始点的不确定性，所以先用 Canopy 进行粗聚类得到聚类数，这是因为 Canopy 聚类得到聚类簇的所用时间很短，并且抗噪能力强，接着因为最大最小距离积法能够选取密度较大的点，使初始点相隔较远，避免挨得很近过于稠密，所以引入最大最小距离积法确定初始点，接着

用 K-means 算法对数据进行聚类。由于用修正余弦相似度计算相似度时，没有考虑用户属性之间的相似度，所以用融合用户属性的相似度计算，对目标用户未评分项目预测评分，并产生推荐。

基于改进 K-means 聚类和融合用户属性的协同过滤推荐算法的流程图如图 1 所示：

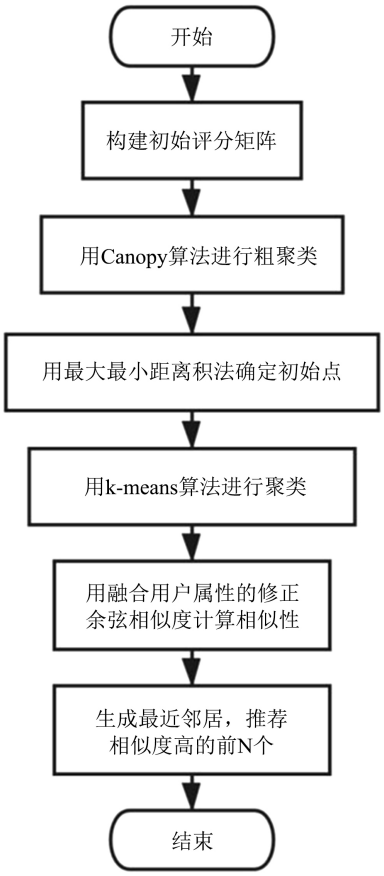

Figure 1. Collaborative filtering recommendation algorithm based on improved K-means clustering and user attribute fusion
图 1. 基于改进 K-means 聚类和融合用户属性的协同过滤推荐算法

4. 实验数据

4.1. 实验数据集

本文使用是由明尼苏达大学 Grouplens 研究项目收集的 Movielens-100 K 数据集如表 1 所示。每个用户至少对 20 部电影进行 1 到 5 的评级。该实验使用 5-折交叉验证方法实验。

Table 1. The data set
表 1. 数据集

类型	用户数量/个	电影数量/个	评分数量/个	稀疏度/%
MovieLens 100 k	943	1682	100,000	93.7
MovieLens 1 M	6040	3883	10,002,209	95.7
MovieLens 10 M	71,567	10,681	10,000,054	98.7

4.2. 评估指标

本文在衡量推荐性能时，采用的是平均绝对误差(MAE)和均方根误差(RMSE)，用来评估提出算法的预测性能[8] [9]，偏差越小推荐质量越高。

计算公式如下：

$$\mathrm{MAE}=\frac{\sum_{u\in T}\left|P_u(t)-P'_u(t)\right|}{|T|} \tag{6}$$

$$\mathrm{RMSE}=\sqrt{\frac{\sum_{u\in T}\left(P_u(t)-P'_u(t)\right)^2}{|T|}} \tag{7}$$

其中，T 为测试集，$|T|$ 为测试集大小，$P_u(t)$ 为用户对项目的预测评分，$P'_u(t)$ 为用户对项目的实际评分。

4.3. 实验结果与分析

4.3.1. 参数 β 的确定

由于新相似度的算法是由修正的余弦相似度与用户属性相似度线性组合得到的，β 的取值对最终计算结果时非常重要的，最终相似度的公式取决于 β 取什么值，β 的取值从 0.1 开始，每次增加 0.1，一直增加到 0.9 停止。观察当 β 取不同值时，平均绝对误差的变化，从而确定 β 的最佳值，进而得到融合用户属性的修正余弦相似度的公式，实验结果如图 2 所示：

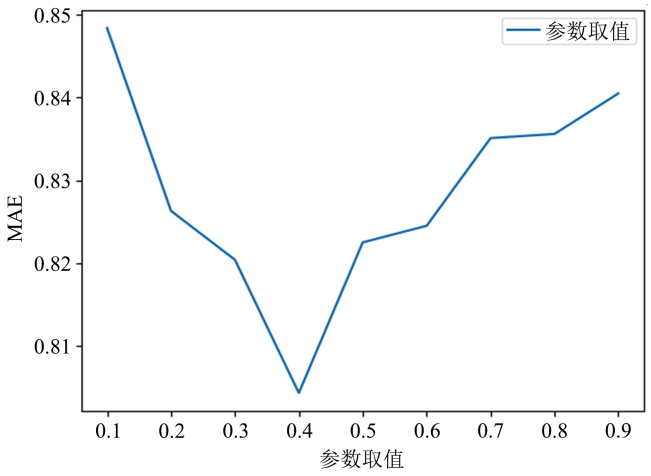

Figure 2. The effect of the parameter β on the MAE value
图 2. 参数 β 对 MAE 值的影响

参数 β 表示混合在一起的修正余弦相似度与用户属性相似度中，用户属性相似度所占的比重，本实验可以看出当 β 取 0.4 的时候，也就是用户属性占四成的时候，平均绝对误差最小，相似度公式的计算效果达到最佳水平。

4.3.2. t_2 的确定

为了得到最佳粗聚类数，图 3 用来验证模型在不同下的准确度情况。

t_2 在[5, 40]区间内，以间隔为 5 来观察聚类数量对实验的影响，结果如图 3 所示在 t_2 为 25 的时候达到最佳，在 t_2 小于 25 时，MAE 的值呈现下降趋势，在 t_2 大于 25 时，MAE 值又呈现缓慢上升的趋势，说明在 t_2 为 25 时，为本文算法模型的最佳状态。

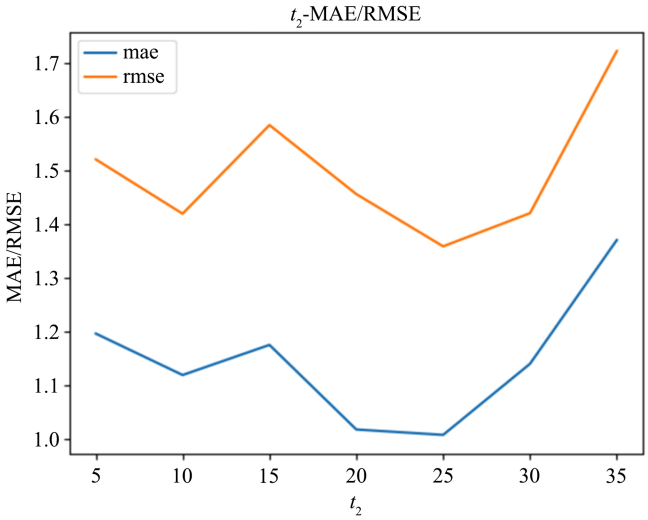

Figure 3. t_2 Impact on MAE and RMSE
图 3. t_2 对 MAE 和 RMSE 的影响

4.3.3. 本文选择以下五种算法进行 MAE、RMSE、两个指标的比较

1) K-means User Clu (KUC)。单侧用户聚类算法，利用 K-means 聚类算法对用户进行聚类。

2) K-means Item Clu (KIC)。利用 K-means 算法对物品进行聚类。

3) Co Clust (CoC)。双侧用户聚类算法，将用户和项目进行联合聚类。

4) Canopy K-means Clu (CKC)。该算法采用改进后的 Canopy K-means 对用户进行聚类。

本实验的邻居用户数选取范围为 0 到 50，间隔值为 10，观察 MAE，RSME 的数值变化。图 4 是五种不同推荐算法在同一数据集上测试的 MAE 指标对比图，图 5 是五种不同推荐算法在同一数据集上测试的 RMSE 指标对比图。

通过图 4 和图 5 可以看出，所有算法的平均绝对误差值，随着最近邻居个数的逐渐增加而逐渐减少，接着又逐步趋于平缓。通过观察可以发现，当最近邻居值在 15 和 45 之间，本文算法能有效改善推荐算法的预测效果，对用户进行聚类能有效地提高推荐质量和推荐精度。

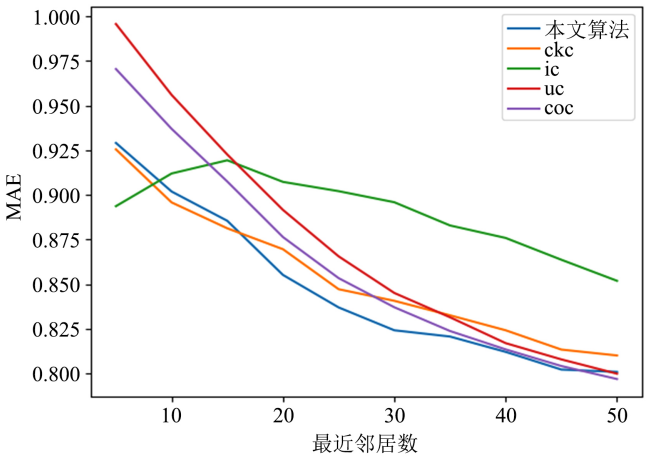

Figure 4. The influence of the number of nearest neighbors on MAE of this algorithm and other algorithms
图 4. 最近邻个数对本文算法与其他算法 MAE 影响

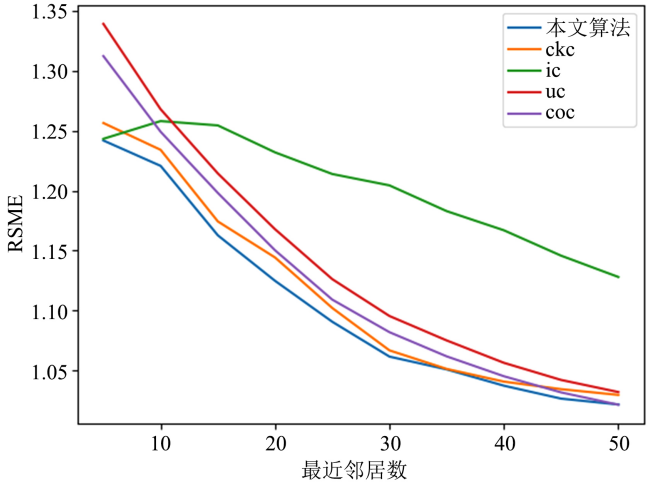

Figure 5. Influence of the number of nearest neighbors on RMSE of the proposed algorithm and other algorithms
图 5. 最近邻个数对本文算法与其他算法 RMSE 影响

5. 结论

本文提出了改进 K-means 聚类与用户属性相似性的协同过滤算法。将最大最小距离积法与混合聚类进行融合，同时用修正余弦相似度与用户属性相结合形成的新相似度算法计算相似性，最后进行相应的推荐。实验表明，无论在 MAE 指标还是在 RMSE 指标的比较上都获得了一定的优势，提出的算法能有效提高推荐算法的预测准确性，且有一定的实际意义。

参考文献

[1] 梁丽君. 基于用户属性聚类的协同过滤推荐算法研究[D]: [硕士学位论文]. 淄博: 山东理工大学, 2018.

[2] 高祥. 基于粗糙聚类的社会化推荐算法研究[D]: [硕士学位论文]. 沈阳: 东北大学, 2016.

[3] 唐泽坤, 黄柄清, 李廉. 基于改进 Canopy 聚类的协同过滤推荐算法[J]. 计算机应用研究, 2020, 37(9): 2615-2619, 2639.

[4] 喻金平, 郑杰, 梅宏标. 基于改进人工蜂群算法的 K 均值聚类算法[J]. 计算机应用, 2014, 34(4): 1065-1069, 1088.

[5] 向小东, 邱梓咸. 基于相似度优化偏差计算的 slope-one 算法研究[J]. 统计与决策, 2019, 35(17): 14-18.

[6] 汪晶. 基于聚类的协同过滤推荐算法研究[D]: [硕士学位论文]. 武汉: 长江大学, 2019.

[7] 赵伟, 林楠, 韩英, 等. 一种改进的 K-means 聚类的协同过滤算法[J]. 安徽大学学报(自科版), 2016(40): 32-36.

[8] Zhang, F., Gong, T., Lee, V.E., *et al*. (2016) Fast Algorithms to Evaluate Collaborative Filtering Recommender Systems. *Knowledge-Based Systems*, **96**, 96-103. https://doi.org/10.1016/j.knosys.2015.12.025

[9] 柳金山. 基于用户动态行为的协同过滤推荐算法研究[D]: [硕士学位论文]. 武汉: 华中科技大学, 2015.