Operations Research and Fuzziology 运筹与模糊学, 2024, 14(5), 400-407
Published Online October 2024 in Hans. https://www.hanspub.org/journal/orf
https://doi.org/10.12677/orf.2024.145481

Hans汉斯

GraphSAGE与遗传算法：推荐系统中的聚类性能优化

代梦飞

上海理工大学管理学院，上海

收稿日期：2024年9月11日；录用日期：2024年10月10日；发布日期：2024年10月18日

摘　要

在信息过载的背景下，如何提高推荐系统的推荐质量和准确性成为了研究热点，而图神经网络作为一种新兴的技术，在这方面展现出了巨大的潜力。本研究首先将四种广泛使用的图神经网络模型GAT、GCN、GIN和GraphSAGE应用于推荐系统中，并通过比较发现GraphSAGE在多项指标上的表现最优。具体来说，GraphSAGE在推荐系统上实现了RMSE (0.6459)、MAE (0.4989)、AUC (0.8701)、Recall (0.6377)、Precision (0.8348)和F1-score (0.7231)的优异性能。进一步，本研究将GraphSAGE与传统的协同过滤算法在推荐系统上的表现进行比较，结果表明GraphSAGE在推荐系统上仍保持优势。在此基础上，本研究还探索了GraphSAGE与遗传算法优化K-means的结合算法——GraphKGA算法对于推荐系统性能的影响，实验证明，GraphKGA算法应用于推荐系统中的RMSE (0.6364)和MAE (0.4855)均取得优异结果，表明其能有效提高推荐系统的性能。

关键词

GraphSAGE模型，遗传算法优化，K-Means聚类，GraphKGA算法，推荐系统

GraphSAGE and Genetic Algorithms: Clustering Performance Optimization in Recommender Systems

Mengfei Dai

Business School, University of Shanghai for Science and Technology, Shanghai

Received: Sep. 11th, 2024; accepted: Oct. 10th, 2024; published: Oct. 18th, 2024

文章引用: 代梦飞. GraphSAGE 与遗传算法: 推荐系统中的聚类性能优化[J]. 运筹与模糊学, 2024, 14(5): 400-407.
DOI: 10.12677/orf.2024.145481

Abstract

In the context of information overload, how to improve the quality and accuracy of recommendation systems has become a research hotspot, and graph neural networks, as an emerging technology, have shown great potential in this regard. This study first applied four widely used graph neural network models, GAT, GCN, GIN, and GraphSAGE, to the recommendation system, and found through comparison that GraphSAGE performed best in multiple indicators. Specifically, GraphSAGE achieved excellent performance in RMSE (0.6459), MAE (0.4989), AUC (0.8701), Recall (0.6377), Precision (0.8348), and F1-score (0.7231) in the recommendation system. Furthermore, this study compared the performance of GraphSAGE with traditional collaborative filtering algorithms in the recommendation system, and the results showed that GraphSAGE still maintained its advantage in the recommendation system. On this basis, this study also explored the impact of the GraphKGA algorithm, which is a combination of GraphSAGE and genetic algorithm optimized K-means, on the performance of the recommendation system. Experiments have shown that the RMSE (0.6364) and MAE (0.4855) of the GraphKGA algorithm applied to the recommendation system have achieved excellent results, indicating that it can effectively improve the performance of the recommendation system.

Keywords

GraphSAGE Model, Genetic Algorithm Optimization, K-Means Clustering, GraphKGA Algorithm, Recommendation System

1. 引言

推荐系统旨在通过分析用户的偏好来预测他们的潜在需求，并提供个性化的推荐服务。用户的偏好可以通过显式或隐式反馈数据推断出来，但大多数现有的协同过滤方法依赖于显式反馈数据，在评分数据稀疏时表现并不理想[1]。因此，需要采取新的技术来提高推荐系统的性能。而近年来，图神经网络因其处理图结构数据方面的优势而广受关注，被广泛地应用于推荐系统中，其中 GAT、GCN、GIN 和 GraphSAGE 等模型显示出了巨大的潜力。

图神经网络的优势在于它提供了强大而系统的工具来探索在推荐系统中被证明有益的多跳关系[2]。它能够通过图结构捕捉用户和物品之间复杂的、非直接的联系，并提供了一套系统化的方法来建模和分析这些关系，从而提高推荐系统的性能。凭借其优势，在推荐系统方面已取得显著的成功和很多先进的成果。Rianne van den Berg 等[3]人提出了外部信息与交互数据相结合的方法——图卷积矩阵补全算法，算是较早地将图卷积应用于推荐系统的工作，可以缓解冷启动问题。杨宇森[4]针对传统协同过滤算法忽视用户与项目关系的问题，提出了一种改进的 LGCN-A 模型，实验得出该模型在多个数据集上是有效性的。吕艳霞等[5]人提出了 SRGN——一种将整合用户的社交关系和物品间的语义联系到算法架构中的社会化推荐算法，结果显示提高了推荐的准确性。张安勤[6]等人提出了一种捕捉用户的动态偏好且考虑社交关系对用户偏好影响的社会感知顺序推荐模型 GASR，实验表明该模型在推荐性能上比现有模型更优。多项研究表明，图神经网络通过捕捉用户和物品之间的复杂关系，提高了推

荐系统的准确性和鲁棒性。

K-means 算法虽然是推荐系统中用户和项目聚类任务中的常用方法，但该算法对初始聚类中心的选择很敏感，就可能会导致不稳定的聚类结果。为解决这个问题，很多研究将遗传算法与 K-means 算法进行结合，均取得了显著的成果。刘婷等[7]提出了一种融合遗传算法的 K-means 聚类方法，旨在通过结合遗传算法的全局搜索能力来对 K-means 算法的初始聚类中心选择进行优化，以此提高聚类的准确性和稳定性。王敞等[8]通过遗传算法操作来优化聚类过程，减少了局部最优解从而提高聚类质量。吕强等[9]提出了一种混合遗传算法，结合混沌优化和遗传算法以克服早熟收敛问题并找到最优聚类中心。戴文华等[10]通过结合并行处理和遗传算法，提高了聚类的全局优化能力和效率，特别在大规模数据处理中表现优异。另外，赖玉霞等[11]提出了一种结合遗传算法和 K-means 算法的聚类方法，有效解决了 K-means 算法对初始聚类中心敏感和容易陷入局部最优的问题，从而提高了聚类的稳定性和准确性。遗传算法和 K-means 的结合取得这样优异的效果，是因为遗传算法的全局搜索能力能够有效地优化初始聚类中心的选择，从而减少对初始条件的依赖，而聚类的高稳定性和准确性就意味着在推荐系统中能够更准确地对用户和物品分类，从而提高推荐结果的准确性。

虽然遗传算法和 K-means 算法的结合在初始聚类中心选择和提高聚类稳定性方面已取得显著成效，但在面对处理大规模图数据以及动态变化的用户偏好时，仍存在一定的局限性。图神经网络，特别是 GraphSAGE，擅长捕捉节点的局部邻域结构，并且能够生成高质量的节点嵌入表示，这些嵌入可以用于增强聚类算法的性能，从而提高推荐系统的准确性。因此，本文基于赖玉霞等人提出的基于遗传算法的 K 均值聚类算法，将其与 GraphSAGE 模型相结合，形成了 GraphKGA 算法，并通过实验来验证 GraphKGA 算法是否能够提高推荐系统的性能。

2. 相关理论

2.1. GraphSAGE 模型

图神经网络是一类专门用来处理图结构数据的深度学习模型，它可以保留图结构信息，并同时对节点和边进行特征学习和表示学习。其核心思想是将来自邻居节点的特征信息进行迭代聚合，并融合在传播过程中聚合的邻居信息与当前中心节点的特征表示。图神经网络可堆叠多个传播层，这些层由聚合操作(Aggregation)和更新操作(Update)组成[12]，聚合操作负责从每个节点的邻居节点收集信息并通过某种方式将邻居节点的特征聚合起来，得到一个综合的邻居节点表示；更新操作将聚合后的邻居信息与当前节点自身的信息结合起来，更新当前节点的表示。可以表示为：

$$\begin{aligned}&\text{Aggregation}: m_v^{(l)} = \text{Aggregator}_l\left(\left\{d_u^{(l)}, \forall u \in M_V\right\}\right)\\&\text{Updata}: d_v^{(l+1)} = \text{Updater}_l\left(d_v^{(l)}, m_v^{(l)}\right)\end{aligned} \tag{1}$$

其中，$d_u^{(l)}$ 代表节点 u 在 L 层的表示，Aggregator_l 和 Updater_l 分别表示第 L 层聚合操作和更新操作的函数。

GraphSAGE 模型是一种灵活的图神经网络，其通过采样和聚合邻居节点信息，实现了在大规模图数据上的高效训练。与传统 GNN 的学习方法不同，GraphSAGE 并不需要为每一个节点训练单独的嵌入表示，而是通过学习一个可以先从目标节点的邻域进行取样，然后再聚合特征的函数。在 GraphSAGE 模型中，首先对每个节点的邻居进行采样，将采样得到的邻居节点特征进行聚合，并通过一个函数来更新节点自身的嵌入表示。在图数据中，它被广泛地应用于节点分类、推荐系统等任务，特别适用于处理大规模和动态图数据。

2.2. 遗传算法和 K-Means 算法

2.2.1. 遗传算法

遗传算法是一种随机化的搜索方法，主要通过模拟自然界的遗传进化过程搜索出最优解。遗传算法不依赖于搜索空间的先验知识或者辅助信息，而是直接根据适应度函数的值来对解的质量进行评估，并且不受函数连续可微和可导的限制，能自由定义搜索范围，通过概率选择指导搜索方向，有强大更优的全局搜索能力。可用下述规划模型来表示：

$$\begin{cases} \max g(x) & (1) \\ x \in Q & (2) \\ Q \subset H & (3) \end{cases} \tag{2}$$

其中，(1)为目标函数，x 为决策变量，(2)和(3)均为约束条件，Q 为基本空间 H 的子集。X 为满足约束条件的可行解，R 为组成的集合即可行解集合[13]。

2.2.2. K-Means 算法

K-means 算法最早由 Mac [14]提出，因其原理简单且效率高而被广泛使用。聚类算法的核心思想是：先选择 k 初始聚类中心，这里的 K 是指定的聚类数目。其次，计算每个数据点与所有初始聚类中心之间的距离，随后根据最小距离原则，将每个数据点归入相应的聚类中心所代表的集合，形成初始簇。接着，基于每个簇内所有数据点的特征，重新计算并更新每个簇的中心位置。最后，重复此步骤，直到簇成员不再变化或目标函数满足条件，最终得到稳定的聚类结果。

3. 方法

本文使用了公开的 movielens100k 真实数据集(http://www.grouplens.org)，该数据集包含 943 位用户对 1682 部电影的 100,000 个评分，评分为 1 到 5，每个用户都至少对 20 部 5 星电影进行了评分。为构建推荐系统，首先对原始数据进行预处理，构建用户—物品评分矩阵，后使用 KNN 算法对评分矩阵中的缺失值进行了填补。

3.1. 图神经网络在推荐系统中的应用

本研究使用了四种不同的图神经网络模型 GAT、GCN、GIN 和 GraphSAGE 对推荐模型进行建模，并比较其性能。首先，构建用户—物品图，通过图数据结构连接用户与物品，并使用图神经网络将用户和物品的关系映射到高维空间中的点，以精准捕捉它们之间的交互关系。其次，对于 GraphSAGE 模型，将其作为推荐系统的核心模型，来对用户和物品的特征进行学习，生成用户和物品的高维嵌入表示。最后，在训练完成后，使用了评估指标 RMSE，MAE 和 AUC 来衡量不同模型的性能，并进一步使用评估指标 Recall，Precision 和 F1-score 来对 GraphSAGE 在推荐系统中的应用效果进行补充评价。

另外，为更进一步确认 GraphSAGE 模型在推荐系统中的性能，将其与传统的协同过滤算法(基于用户的协同过滤和基于物品的协同过滤)进行比较，评估指标为 RMSE 和 MAE。

3.2. GraphSAGE 与遗传算法优化 K-Means 的结合(GraphKGA 算法)

在前人的基础上，本文将 GraphSAGE 与遗传算法优化 K-means 进行结合，形成了 GraphKGA 算法，并通过评估指标 RMSE 和 MAE 来验证是否会提高推荐系统的性能。首先，使用 GraphSAGE 模型生成用户和物品的高维嵌入向量。其次，使用遗传算法优化 K-means 的聚类过程，以找到最优的聚类中心，最终达到提高推荐系统性能的目的。GraphKGA 算法融合了 GraphSAGE 的强大嵌入能力和遗传算法的全局

搜索能力，最终通过优化的聚类中心来为用户提供最相关的物品。

3.3. 评估指标

本文实验主要使用的评估指标是 RMSE，MAE 和 AUC。

(1) 均方根误差(RMSE)

$$\mathrm{RMSE}=\sqrt{\frac{1}{N}\sum_{i=1}^{n}\left(y_i-\hat{y}_i\right)^2} \tag{3}$$

RMSE 用于衡量模型预测值与实际值间的差异，它能够反映预测误差的平方平均值，从而更好地突出较大的误差。RMSE 的数值越小，说明模型的预测效果越好。

(2) 平均绝对误差(MAE)

$$\mathrm{MAE}=\frac{1}{n}\sum_{i=1}^{n}\left|y_i-\hat{y}_i\right| \tag{4}$$

MAE 用于衡量预测值与实际值之间的平均绝对差异。MAE 的数值越小，说明模型的预测效果越好。

(3) ROC 曲线下面积(AUC)

$$\mathrm{AUC}=\int_0^1 \mathrm{TPR}\left(\mathrm{FPR}\right)\mathrm{dFPR} \tag{5}$$

其中，TPR 和 FPR 分别为真阳率和假阳率。

AUC 用于衡量分类器的识别性能，数值范围 0~1。AUC 的数值越接近 1，说明分类器的效果越好。

4. 实验结果和分析

4.1. 四种图神经网络模型

4.1.1. 各模型的损失历史(Loss History)

图 1 显示了 GAT、GCN、GIN、GraphSAGE 模型在训练过程中迭代 100 次的损失值记录。

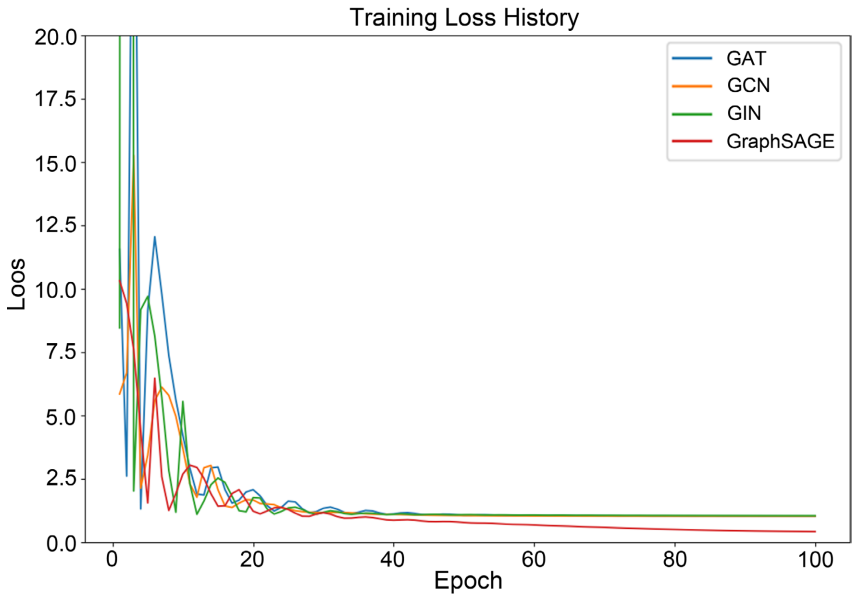

Figure 1. Loss values of GAT, GCN, GIN, and GraphSAGE models after 100 iterations of training
图 1. GAT、GCN、GIN、GraphSAGE 模型在训练过程中迭代 100 次的损失值

在图 1 中可以看到四种模型在训练过程中整体上均呈损失值逐渐下降的趋势，这是因为随着训练的进行，模型的预测性能持续地提升。虽然四个模型整体趋势一样，但损失值的波动情况是不相同的。其中，在训练初期，GraphSAGE 模型显示出了最快的收敛速度，其损失值从最开始就快速下降，并且在较早期的 Epoch 中就达到了较低的水平，这表明其学习能力较强。在训练中后期，GraphSAGE 模型显示出相对较小的损失波动，这表明其预测更加稳定。在训练结束时，GraphSAGE 的损失值最低，这意味着 GraphSAGE 性能相对较好。而 GAT、GCN 和 GCN 在初期的波动较为剧烈，且损失值趋势均大于 GraphSAGE 模型。综合来说，GraphSAGE 模型在训练过程中表现出了更快的收敛速度、更小的损失波动以及更低的最终损失值，这表明在这四个图神经网络模型中其性能最优。

4.1.2. 各指标值比较

图 2 显示了 GAT、GCN、GIN、GraphSAGE 模型在 Movielens100k 数据集上的 RMSE，MAE 和 AUC 结果。

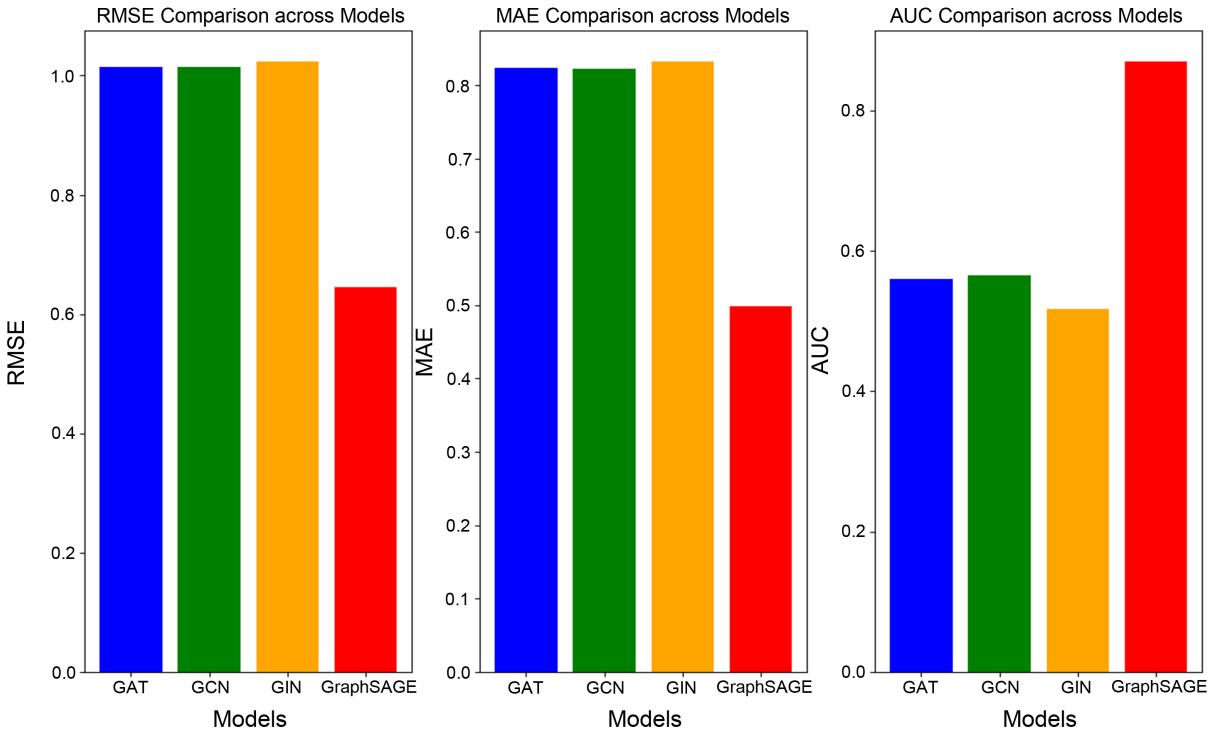

Figure 2. Three indicator values of GAT, GCN, GIN, and GraphSAGE models on the Movielens100k dataset
图 2. Movielens100k 数据集上 GAT、GCN、GIN、GraphSAGE 模型的三项指标值

(1) RMSE

GraphSAGE 模型在 RMSE 指标上表现最佳(0.6459)，表示其预测的评分与实际评分的差异最小。GCN 模型(1.0138)、GAT 模型(1.0150)和 GIN 模型(1.0237)次之，表现不好。

(2) MAE

GraphSAGE 模型在 MAE 指标上表现最好(0.4989)，这进一步验证了其在预测评分方面的准确性优于其他三个模型。而 GAT 模型(0.8241)、GCN 模型(0.8229)和 GIN 模型(0.8327)的 MAE 值较高，表现较差。

(3) AUC

GraphSAGE 模型在 AUC 指标上表现最为优(0.8701)，表明其在区分正负样本(本文指高评分和低评

分物品)方面效果最佳。而 GAT 模型(0.5605)、GCN 模型(0.5654)和 GIN 模型(0.5174)的 AUC 值相对较低，说明其在分类能力上较差。

综合来看，GraphSAGE 模型在三个评估指标上均表现为优异，这表明 GraphSAGE 模型在推荐任务中具有较高的预测准确性及分类能力。同时，为补充验证 GraphSAGE 模型在推荐系统中的应用性能，又对 GraphSAGE 模型添加了三个评估指标来做补充，即 Recall (0.6377)、Precision (0.8348)和 F1-score (0.7231)，结果证明，GraphSAGE 模型在补充评估指标上同样表现突出，这进一步证明其在预测用户偏好方面表现良好，能为用户提供更高质量的推荐结果。因此得出，GraphSAGE 模型在进行推荐任务时，相比于其他三个图神经网络模型，推荐质量更高。

4.2. GraphSAGE 模型和传统协同过滤算法

为进一步确认 GraphSAGE 模型在推荐系统中的性能，将其与传统的协同过滤算法(基于用户的协同过滤和基于物品的协同过滤)进行了比较，图 3 显示了 GraphSAGE 模型、基于用户的协同过滤算法和基于物品的协同过滤算法在 RMSE 和 MAE 指标上的表现。

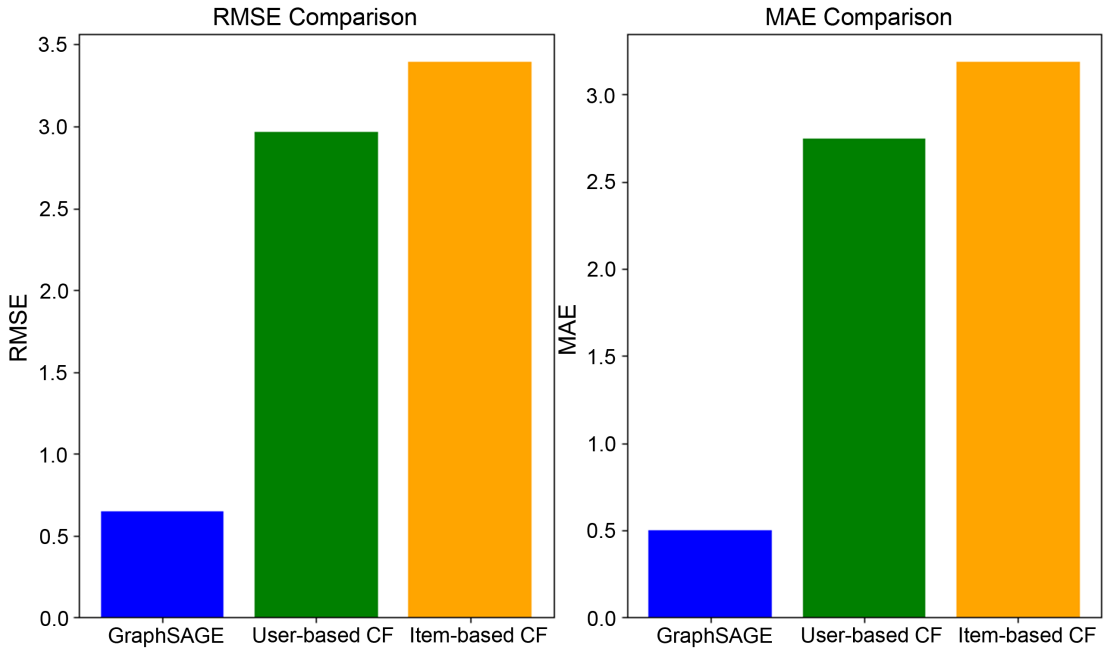

Figure 3. RMSE and MAE values of the GraphSAGE model, user-based collaborative filtering algorithm, and item-based collaborative filtering algorithm
图 3. GraphSAGE 模型、基于用户的协同过滤算法和基于物品的协同过滤算法的 RMSE 和 MAE 指标值

GraphSAGE 模型的表现显著优于协同过滤模型：GraphSAGE 模型的 RMSE (0.6459)和 MAE (0.4989)均远小于基于用户和基于物品的协同过滤算法的 RMSE 和 MAE。另外，也发现其他三种图神经网络模型的 RMSE 和 MAE 远小于基于用户和基于物品的协同过滤算法。这表明，图神经网络模型，尤其是 GraphSAGE 模型在推荐系统中的应用效果比传统协同过滤算法更优。

对于 GraphSAGE 模型和传统协同过滤算法的效果差异分析，这可能是由于 GraphSAGE 模型将用户和物品看作是图中的节点，然后利用图卷积层捕捉节点间的复杂关系，更好地学习到了用户和物品之间的潜在特征，从而提高推荐效果。而相比之下，传统的协同过滤模型是仅仅依赖用户—物品评分矩阵的相似度，无法充分地捕捉用户和物品间复杂的交互信息，因此表现较差。

4.3. GraphKGA 算法

在本次实验中，GraphKGA 算法经过训练和评估，最终达到了 RMSE (0.6364)和 MAE (0.4855)的性能指标，其中，RMSE 和 MAE 的值表明 GraphKGA 算法在预测用户评分时的误差较小，这就意味 GraphKGA 算法具有较高的预测准确性，能够较好地捕捉用户和物品之间的复杂关系，提供更精确的推荐结果。同时，这也进一步证实了结合图神经网络和遗传算法优化聚类的方法在推荐系统领域的应用潜力。

5. 实验结果和分析

本文首先研究了四种图神经网络模型(GAT、GCN、GIN、GraphSAGE)在推荐系统中的应用性能，并与传统的协同过滤推荐算法进行比较。研究表明，基于图神经网络的推荐系统表现出了显著的优势，尤其是 GraphSAGE 模型，在各项指标上均优于另外三个图神经网络模型和传统的协同过滤方法，其各评估指标结果为 RMSE (0.6459)、MAE (0.4989)、AUC (0.8701)、Recall (0.6377)、Precision (0.8348)和 F1-score (0.7231)，并且与传统的协同过滤推荐算法相比，基于图神经网络的推荐算法获得了更好的性能。其次，在前人的基础上提出并验证了 GraphKGA 算法，其 RMSE (0.6364)和 MAE (0.4855)均取得优异的结果，这表明了 GraphKGA 模型具有较高的预测准确性，能够使推荐系统提供更精确的推荐结果。未来工作将考虑应用更大规模的数据集和探索更复杂的图神经网络模型，以进一步提高推荐系统的性能。

参考文献

[1] Feng, J., Xia, Z., Feng, X. and Peng, J. (2021) RBPR: A Hybrid Model for the New User Cold Start Problem in Recommender Systems. *Knowledge-Based Systems*, **214**, Article 106732. https://doi.org/10.1016/j.knosys.2020.106732

[2] Wu, S., Sun, F., Zhang, W., Xie, X. and Cui, B. (2022) Graph Neural Networks in Recommender Systems: A Survey. *ACM Computing Surveys*, **55**, 1-37. https://doi.org/10.1145/3535101

[3] Berg, R.V.D., Kipf, T.N. and Welling, M. (2017) Graph Convolutional Matrix Completion. https://www.kdd.org/kdd2018/files/deep-learning-day/DLDay18_paper_32.pdf

[4] 杨宇森. 融合图神经网络的协同过滤算法的改进研究[D]: [硕士学位论文]. 大连: 大连交通大学.

[5] 吕艳霞, 郝帅, 乔广通, 等. 一种基于图神经网络的社会化推荐算法[J]. 东北大学学报(自然科学版), 2024, 45(1): 10-17.

[6] 张安勤, 李然, 田秀霞. 基于图神经网络的社会感知顺序推荐模型[J]. 计算机应用与软件, 2024, 41(3): 246-252+282.

[7] 刘婷, 郭海湘, 诸克军, 等. 一种改进的遗传 k-means 聚类算法[J]. 数学的实践与认识, 2007, 37(8): 104-111.

[8] 王敞, 陈增强, 袁著祉. 基于遗传算法的 K 均值聚类分析[J]. 计算机科学, 2003, 30(2): 163-164.

[9] 吕强, 俞金寿. 基于混合遗传算法的 K-Means 最优聚类算法[J]. 华东理工大学学报(自然科学版), 2005, 31(2): 219-222.

[10] 戴文华, 焦翠珍, 何婷婷. 基于并行遗传算法的 K-means 聚类研究[J]. 计算机科学, 2008, 35(6): 171-174.

[11] 赖玉霞, 刘建平, 杨国兴. 基于遗传算法的 K 均值聚类分析[J]. 计算机工程, 2008, 34(20): 200-202.

[12] 刘天航, 杨晓雪, 周慧, 等. 基于图神经网络的协同过滤推荐算法综述[J]. 集成技术, 2024, 13(4): 1-15.

[13] 金玲, 刘晓丽, 李鹏飞, 等. 遗传算法综述[J]. 科学中国人, 2015(9X): 230.

[14] MacQueen, J. (1967) Some Methods for Classification and Analysis of Multivariate Observations. In: *Berkeley Symposium on Mathematical Statistics and Probability*, University of California Press, 1-15.