基于改进二阶因子分解机算法的地点轨迹预测

邓楷馨[1]，202019040229

（1.成都理工大学 产业技术学院，四川 宜宾 644000）

摘　要：本文提出了一种基于传统的因子分解机（Factorization Machine）模型的改进算法。主要创新在于引入地理空间数据处理和注意力机制。新算法通过聚类技术处理场地地理位置（经纬度数据），增强了模型对地理空间特征的处理能力。引入注意力机制以增强模型对关键特征的关注度，以提高预测的准确性。本文方法在 TKY 数据集上的精确率、召回率、平均精度评分（MAP）和归一化折扣累积增益（nDCG）上均有显著提升，分别提高了 6.4%、4.5%，6.8%、6.2%。本文的实施代码在 https://github.com/dkx2077/SOFM.git。

关键词：因子分解机；地理空间数据；注意力机制；推荐系统

Location Trajectory Prediction Based on Improved Second-Order Factorization Machine Algorithm

Kaixin Deng1, 202019040229

(1. School of Industrial Technology, Chengdu University of Technology, Yibin, Sichuan 644000, China)

Abstract： This paper proposes an improved algorithm based on the traditional Factorization Machine (FM) model, with key innovations including the integration of geospatial data processing and attention mechanisms. Firstly, the new algorithm enhances the model's ability to handle geospatial features by clustering venue locations (latitude and longitude data). Secondly, the introduction of an attention mechanism heightens the focus on key features, thereby improving prediction Precesion. The method presented in this paper significantly improves Precesion, recall, mean average precision (MAP), and normalized discounted cumulative gain (nDCG) on the TKY dataset, with increases of 6.4%, 4.5%, 6.8%, and 6.2%, respectively. Our implementation code is available at https://github.com/dkx2077/SOFM.git.

Keywords: Factorization Machine; Geospatial Data; Attention Mechanism; Recommendation System.

近年来近年来，因子分解机(Factorization Machine, FM)模型已被广泛应用于多个领域[1-4]，如推荐系统和数据分析。然而，传统的 FM 模型面临着处理高维稀疏数据和地理空间信息的挑战。为此，本文提出了一种新的改进算法，结合了地理空间数据处理和注意力机制，以增强模型的表现力和适应性来预测每周二晚上用户最有可能前往场所。

传统 FM 模型[5,6]主要通过线性关系来预测结果，但在处理高维稀疏数据时，尤其是在地理空间数据应用中，这种方法的效率和准确性并不理想。为了克服这些限制，新算法首先利用聚类技术对地理位置（即经纬度数据）进行处理，将地理实体根据位置进行有效聚类。这一步骤不仅提升了模型对于地理空间特征的理解，而且增强了模型在处理类似于场所推荐这类问题时的准确性和实用性。

其次，新算法引入了注意力机制。在传统 FM 模型中，所有特征被平等对待[7-9]，但实际上并非所有特征对预测结果都同等重要。注意力机制使得模型能够关注于那些对预测结果更有影响力的特征，从而提高了整体的预测准确度。特别是在处理复杂的用户行为数据时，这种方法能更有效地捕捉关键信息。

此外，新算法在保留原有 FM 模型优势的同时，通过整合地理空间信息和注意力机制，显著提高了处理高维稀疏数据的能力。这一改进对于推荐系统尤为重要，因为它能够更准确地理解和预测用户的行为和偏好。

在实验部分，本文使用了公开数据集 FourSquare - NYC and Tokyo Check-ins 来验证新算法的有效性。实验结果表明，与传统 FM 模型相比，新算法在各项性能指标上都有显著提升，特别是在准确度、召回率和平均精确度上。

综上所述，本文的主要贡献在于提出了一种结合地理空间数据处理和注意力机制的改进 FM 模型。该模型不仅在理论上对传统 FM 模型进行了重要的扩展，而且在实际应用中表现出更高的准确性和适用性，尤其是在处理与地理位置相关的复杂数据时。这些改进为推荐系统和数据分析等领域提供了新的视角和方法[10-11]。

1 相关工作

在本节中，先简要回顾了先前关于推荐系统的定义。然后回顾了协同过滤方法。最后，对本文提出的方法与相关工作进行了比较总结。

1.1 推荐系统

推荐系统是一种有助于用户快速发现有用信息的工具，同时也是一种能增加公司产品与用户接触、购买等行为概率的工具。对于用户而言，推荐系统在用户需求不十分明确的情况下，通过利用各类历史信息来猜测其可能喜欢的内容，提供了更精准的信息过滤，相较于传统搜索系统更具优势。对于公司而言，推荐系统的应用有助于解决产品吸引用户、提高用户留存率、增加用户黏性以及提高用户转化率的问题，从而实现连续增长的商业目标。因此，推荐系统主要服务于那些希望提供个性化内容以吸引和满足用户需求的公司，以及那些希望提高产品与用户互动和购买行为概率的公司，从而满足用户需求并促进业务增长的个体和组织。

在推荐系统评测中，常用的评测指标有用户满意度和预测准确度。用户满意度是关键指标之一，直接影响推荐系统的综合性能。它通常需要通过用户调查或在线实验来获取，比如购买率、点击率、用户停留时间和转化率等，用以度量用户对推荐内容的满意程度。预测准确度则是一种离线评测方式，用于衡量推荐系统的预测能力，包括评分预测和 Top-N 预测两方面，前者关注用户是否会给予高评分，而后者关注用户是否会查看所推荐内容的信息。这些评测指标在不同应用场景中具有不同的重要性，但通常是评估推荐系统性能的关键因素之一。

1.2 协同过滤算法

协同过滤算法是推荐系统中最早诞生的算法之一，也是应用最广泛的推荐算法之一。其基本原理是通过寻找与目标用户兴趣相似的其他用户，利用这些相似用户的喜好来推荐物品给目标用户。这一过程包括预测和推荐两个主要步骤。在预测过程中，通过分析目标用户的行为和偏好，构建用户的偏好模型，然后寻找与该模型相似的其他用户。在推荐过程中，将与目标用户相似的用户喜好的物品推荐给目标用户，类似于在日常生活中向朋友咨询电影推荐。协同过滤算法流程如图 1 所示。

协同过滤算法的核心思想是根据用户的历史行为数据，如评价、购买、下载等，为用户推荐物品，而不依赖于物品的附加信息或用户的个人信息。目前，基于邻域的方法是广泛应用的协同过滤算法，其中包括基于用户的协同过滤(UserCF)和基于物品的协同过滤(ItemCF)。这些算法的关键步骤之一是计算用户或物品之间的相似度。

推荐系统的核心部分是推荐算法，不同的算法会影响推荐的准确性。对于协同过滤算法，其流程包括收集用户偏好、找到相似用户和物品，以及形成相似推荐。这些步骤共同构建了推荐系统的核心思想和基本框架。

2 方法与材料

本文提出的算法是一种改进的因子分解机模型，专门用于处理包含地理空间数据的复杂推荐系统问题。其核心在于将传统 FM 模型与地理空间数据处理和注意力机制相结合，以增强模型的性能和适用性。首先，算法通过对用户数据的时间和位置信息进行预处理，包括将时间分解为年、月、日和小时，并将用户 ID、场所 ID、类别等信息进行编码。这一步是为了将这些高维稀疏的类别特征转化为易于模型处理的形式。其次，算法的一个显著创新在于对地理空间数据的处理。它通过 K-Means 聚类算法对场所的地理位置（纬度和经度）进行聚类，形成地理空间特征。这一步不仅增强了模型对地理位置的理解能力，而且使得模型能够更好地在推荐系统中应用，如更准确地推荐用户可能感兴趣的场所。接下来，算法引入了注意力机制，这是对传统 FM 模型的另一大改进。通过计算注意力分数，模型能够专注于对预测结果更重要的特征，从而提高了预测的准确性

和效率。整个模型的训练过程包括多次迭代，每次迭代都会根据交叉熵损失函数更新模型的参数。模型的性能通过准确度、召回率、平均精度评分和归一化折扣累积增益来评估。这些指标共同体现了模型在推荐系统中的有效性[12]，特别是在考虑用户行为和地理位置信息的复杂场景中。本文通过结合传统因子分解机模型、地理空间数据处理和注意力机制，有效地提高了推荐系统的性能，特别是在处理地理位置相关的数据时表现出了显著的优势。

图 1 协同过滤算法的流程

Fig. 1 The flow of collaborative filtering algorithmmaximizing mutual information

2.2 算法框架

本文提出算法的框架可以分为三个主要部分：

1. 数据预处理和特征编码：算法从原始数据中提取并处理时间和地理空间信息。它将时间信息分解为年、月、日和小时，并将用户 ID、场所 ID、类别等信息进行编码，形成易于模型处理的特征向量。

2. 模型结构与训练：核心模型是一个改进的因子分解机，引入了注意力机制。它通过学习输入特征的线性组合和交叉特征的非线性交互来预测结果。注意力机制用于强化对重要特征的关注。模型通过交叉熵损失函数进行训练，优化器采用 Adam，搭配学习率调度器进行训练过程的调整。

3. 性能评估和模型保存：通过准确度、召回率、平均精度评分和归一化折扣累积增益等指标评估模型性能。在达到最佳性能时，模型状态被保存，以便于后续的应用或分析。

整个算法框架结合了机器学习中的多种技术和方法，从数据预处理到模型训练、再到性能评估和模型持久化，形成了一个完整的流程，特别适用于处理高维稀疏数据和地理空间信息相关的推荐系统问题。

2.3 实现

2.3.1 数据预处理和特征编码

时间数据处理：使用`pd.to_datetime`将签到时间（`checkin_times`）转换为 Pandas 的 DateTime 对象，然后提取年、月、日和小时作为单独的特征。

类别特征编码：对于用户 ID、场所 ID、类别等非数值型特征，使用`LabelEncoder`将它们转换为数值型编码，以便模型能够处理。

2.3.2 构建稀疏矩阵

特征转换：将编码后的特征以及从时间数据提取的特征转换为稀疏矩阵。每个特征都被转换为一个 CSR（Compressed Sparse Row）格式的稀疏矩阵。

地理空间数据聚类：使用`KMeans`聚类算法处理纬度和经度数据，将地理位置归类为若干聚类中心，然后将聚类结果也转换为稀疏矩阵。

稀疏矩阵一共包括用户 ID、场所 ID、类别名称、时间的年、月、日和小时等特征以及本文新增的经纬度特征。

2.3.3 注意力机制

构建注意力层：定义一个线性层，用于计算特征的注意力分数。利用`nn.Linear`建立一个从输

入维度到 1 的线性映射，然后通过`softmax`函数计算注意力权重。

2.3.4 因子分解机模型

目标：预测用户每周二最有可能出现场所。

线性部分：使用`nn.Linear`层处理输入特征，计算线性关系。

交互部分：通过定义参数矩阵`v`，计算特征之间的交叉影响。使用矩阵乘法和逐元素乘法来计算二次项。

应用注意力：将注意力权重应用到交互部分的计算中，以增强模型对重要特征的关注。

2.3.5 模型训练

数据加载：使用`DataLoader`来批量加载稀疏矩阵和标签，方便模型训练。

损失函数和优化器：使用交叉熵损失(`nn.CrossEntropyLoss`)和 Adam 优化器(`optim.Adam`)，以及学习率调度器(`torch.optim.lr_scheduler.StepLR`)。

训练循环：在每个 epoch 中，计算模型在训练集上的损失并进行反向传播，调整模型参数。

2.3.5 损失函数

本文选择常用的 infoNCE[17]作为损失函数。

将这两部分损失线性组合后作为模型整体损失。

2.3.6 性能评估

评估指标：使用精确率、召回率、平均精度评分（MAP）和归一化折扣累积增益（nDCG）等指标来评估模型在测试集上的性能。

无监督评估：使用`torch.no_grad()`在模型评估阶段关闭梯度计算，以减少内存消耗并加速计算过程。

2.3.7 模型保存和记录

保存最优模型：在性能达到最佳时，使用`torch.save`保存模型状态。

结果记录：将每个 epoch 的损失和评估指标记录下来，并使用`pandas.DataFrame`以及`to_excel`方法保存为 Excel 文件，方便后续分析。

2.4 算法伪代码

综上所述，最终本文算法的整个流程伪代码如算法 1 所示：

算法 1 算法流程

整体流程伪代码

```
输入:
  用户 ID 集 user_ids
  场所 ID 集 venue_ids
  类别集 categories
  签到时间集 checkin_times
  纬度集 latitudes
  经度集 longitudes
  迭代次数 epochs
  数据集 data
输出:
  模型 model
  性能评估 metrics

function preprocess_data(user_ids, venue_ids, categories, checkin_times, latitudes, longitudes):
  对输入数据进行预处理和编码
  返回稀疏矩阵 X

function train_model(X, data, epochs):
  初始化模型和优化器
  对数据集进行迭代训练
  返回训练后的模型

function evaluate_model(model, X_test, y_test):
  使用测试集评估模型性能
  返回评估指标
```

3 实验设置

实验使用广泛使用的 FourSquare - NYC and Tokyo Check-ins 数据集来评估模型的性能，测试指标为精确率、召回率、平均精度评分（MAP）和归一化折扣累积增益（nDCG）。本节将概述所使用的数据集、实验环境细节以及对结果的讨论。代码基于 Pytorch 和 sklearn 库实现，在一个 6GB 显存的 NVIDIA 2060 上执行所有实验。

3.1 数据集

为了评估本文的方法，本文采用了基准数据集 FourSquare - NYC and Tokyo Check-ins[13]数据集。该数据集包含大约 10 个月（从 2012 年 4 月 12 日到 2013 年 2 月 16 日）收集的纽约和东京签到信息。其中包含纽约市的 227,428 次签到和东京的 573,703 次签到。每次签到都与其时间戳、GPS 坐标和语义相关联（由细粒度的场地类别表示）。该数据集最初用于研究 LBSN 中用户活动的时空规律。

3.2 评估方案

所有实验都使用精确率、召回率、平均精度评分（MAP）和归一化折扣累积增益（nDCG）等指标来评估模型在测试集上的性能。

3.3 基线方法

为了比较本文提出的算法与以往的工作，本文选择了和传统的二阶交互因子分解算法对比实验。对于所有基线方法，本文采用了相同的超参数。其中 epoch 为 100，学习率 lr 为 0.01，学习率调度器 gamma 为 0.1，step 为 10，神经网络输入维度 n 为稀疏矩阵的列数，隐藏层维度 k 为 15，聚类算法聚类数量为 10。

4 结果与分析

4.1 实验结果

表 1 在数据集上验证结果

Table 1 Validation results on dataset

Algorithm	Dataset	Precesion	Recall	MAP	nDCG
Original	NYC	**0.173**	0.092	0.199	**0.195**
algorithm	TKY	0.377	0.173	0.419	0.417
OURS	NYC	0.170	**0.105**	**0.201**	**0.195**
	TKY	**0.441**	**0.218**	**0.487**	**0.479**

本文方法在数据集上与基线方法进行了对比。结果如表 1 所示，表中数据为节结果对比，其中加黑的结果为最好表现。本文的算法在分类精确率方面与基线方法相比表现出具有竞争力的性能，在大多数数指标上超越了基线算法，TKY 数据集的对比实验的可视化结果如图 2 所示。

与基准对比方法相比，本文在 TKY 数据集上精确率、召回率、平均精度评分（MAP）和归一化折扣累积增益（nDCG）分别提高了 6.4%、4.5%，6.8%、6.2%，在 NYC 数据集上召回率、平均精度评分（MAP）分别提高了 1.3%、0.2%。在 TKY 数据集上与传统二阶因子分解方法相比，本文提出的算法展现了优越的性能，验证了本文提出算法的有效性。并且本文提出的改进算法拥有收敛更快，稳定的优点。本文提出的算法在 20 个 epoch 附件便能实现收敛，这能够极大程度减少训练成本。

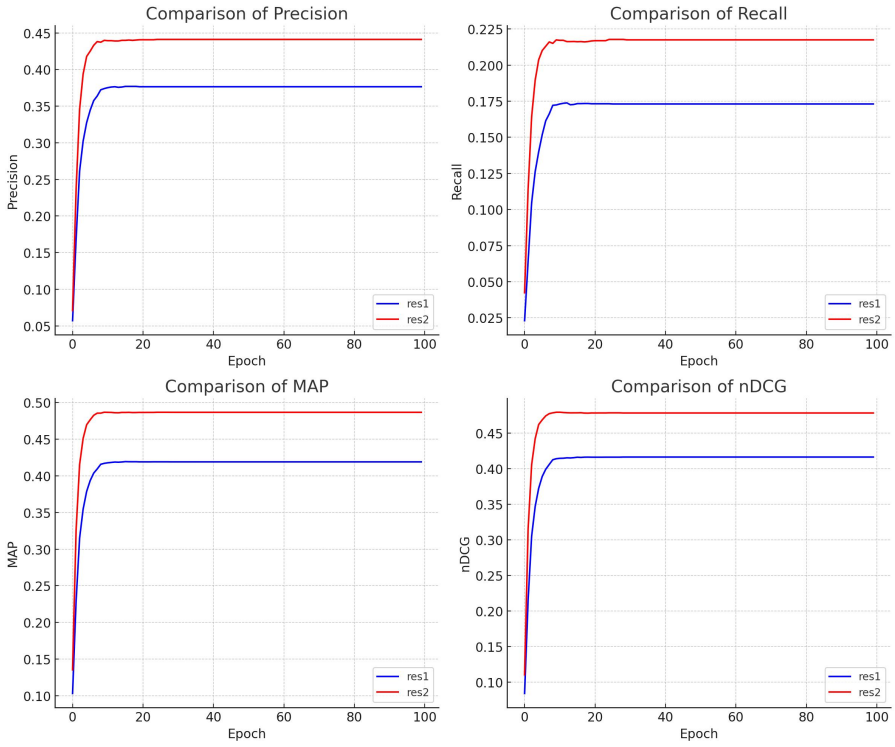

图 2 对比实验结果

Fig. 2 Comparative experimental results

5 总结与结论

本文提出了一种全新的改进的因子分解机模型，通过整合地理空间数据处理和注意力机制，显著提高了推荐系统的性能。在 TKY 和 NYC 数据集上的实验结果充分证明了该算法的有效性和优越性。与传统二阶因子分解方法相比，本文方法在 TKY 数据集上的精确率、召回率、平均精度评分（MAP）和归一化折扣累积增益（nDCG）上均有显著提升，分别提高了 6.4%、4.5%，6.8%、

6.2%。在 NYC 数据集上，召回率和 MAP 也分别提高了 1.3%和 0.2%。这些成果不仅验证了本文算法在理论上的创新性，也展示了其在实际应用中的强大潜力。特别是在处理高维稀疏数据和地理空间信息方面，本文方法展现了明显的优势。注意力机制的引入进一步增强了模型对关键特征的关注，提高了预测的准确性和效率。未来的工作可以在更广泛的应用场景中测试和优化本文提出的算法，如个性化广告推荐和社交网络服务，以进一步证明其适用性和可扩展性。此外，考虑到地理空间数据的多样性，未来研究还可以探索将更多类型的地理空间信息整合到推荐系统中，以进一步提高模型的泛化能力和实用价值。

参考文献:

[1] 余秋宏. 基于因子分解机的社交网络关系推荐研究[D]. 北京: 北京邮电大学, 2013.

[2] 胡亚慧, 李石君, 余伟, 等. 一种结合文化和因子分解机的快速评分预测方法[J]. 南京大学学报: 自然科学版, 2015, 51(4): 826-833.

[3] 杨逍. 机器学习中隐式因子模型及其优化算法研究[D]. 哈尔滨工业大学, 2013.

[4] 孙海峰, 张勇, 解静芳. 正定矩阵因子分解模型在环境中多环芳烃源解析方面的应用[J]. 生态毒理学报, 2015, 10(4): 25-33.

[5] 秦大路, 李晓宇. 基于层次化上下文因式分解机的推荐系统[J]. 河南师范大学学报 (自然科学版), 2015, 2.

[6] 邱煜炎, 吴福生. 基于粒子群优化的因子分解机算法[J]. 山西师范大学学报: 自然科学版, 2020, 34(1): 5-11.

[7] 陈子晋. 基于因子分解机的推荐算法研究[D]. 北京邮电大学, 2021.

[8] 方文剑. 基于用户行为序列的推荐算法研究及应用[D]. 东华大学, 2022.

[9] 刘宁宁. 基于深度因子分解机和循环神经网络的位置预测研究[D]. 重庆邮电大学, 2020.

[10] Sun Y, Pan J, Zhang A, et al. FM2: Field-matrixed factorization machines for recommender systems[C]//Proceedings of the Web Conference 2021. 2021: 2828-2837.

[11] Chen T, Yin H, Nguyen Q V H, et al. Sequence-aware factorization machines for temporal predictive analytics[C]//2020 IEEE 36th International Conference on Data Engineering (ICDE). IEEE, 2020: 1405-1416.

[12] Chen Y, Wang Y, Ren P, et al. Bayesian feature interaction selection for factorization machines[J]. Artificial Intelligence, 2022, 302: 103589.

[13] https://www.kaggle.com/code/rahuljois/finding-a-particular-user

作者简介：邓楷馨，成都理工大学智能科学与技术专业本科生，主要研究方向：深度学习，目标检测，三维重建