2024届研究生硕士学位论文

分类号：_____ 学校代码：10269

密 级：_____ 学 号：51244407096

華東師範大學

East China Normal University

硕士学位论文

MASTER'S DISSERTATION

论文题目：适用于推荐系统的迁移线性模型

院 系：统计学院

专 业 学 位 类 别：应用统计硕士

专 业 学 位 领 域：应用统计

研 究 方 向：大数据统计与人工智能

学 位 申 请 人：马千里

指 导 教 师：张日权 教授

2024 年 5 月

MASTER DISSERTATION 2024 UNIVERSITY CODE：10269

STUDENT NO：51244407096

EAST CHINA NORMAL UNIVERSITY

Title:A Transfer Linear Model for Recommendation Systems

Department: School of Statistics

Category: Master of Applied Statistics

Field: Applied Statistics

Research Focus: Big Data and AI

Candidate: Ma Qianli

Supervisor: Professor Zhang Riquan

May of 2024

马千里硕士学位论文答辩委员会成员名单

姓名	职称	单位	备注
方方	教授	华东师范大学	
陈立峰	研究员	上海富数科技有限公司	
危佳钦	教授	华东师范大学	

摘 要

随着信息规模的不断扩大和技术的不断进步，个性化推荐逐渐成熟和普及。推荐技术不断发展的同时，也出现了许多新的挑战。信息茧房和冷启动是现代推荐系统最常见的问题，它们会导致推荐系统的多样性降低、准确度下降。

为了解决这两个问题，跨域推荐成为了近期的研究热门。跨域推荐利用迁移学习技术提取其他领域丰富的信息，以提高稀疏领域的推荐性能。虽然近年来对跨域推荐以及统计推荐模型进行了广泛的研究，但是到目前为止，还没有看到将统计推荐模型用于跨域推荐的详细研究。为此本文提出了用于推荐系统的迁移线性模型，旨在利用跨域推荐迁移多域信息，并结合线性模型可解释性的优势，大幅缓解信息茧房和冷启动问题。

本文在跨域推荐领域常用的 MovieLens-Netflix、Douban 以及 Amazon 数据集上对所提出算法的有效性进行了检验。实验结果表明：利用源域的信息能够有效提升模型在目标域的效果，所提出的算法能够在多域情形下有效缓解冷启动及信息茧房问题。

关键词： 推荐系统 迁移学习 因子分解机 冷启动 信息茧房

ABSTRACT

With the continuous expansion of information scale and technological advancements, personalized recommendations have gradually matured and become popular. As recommendation technologies continue to evolve, many new challenges have emerged. Information silos and cold start are the most common issues in modern recommendation systems, leading to decreased diversity.

To address these issues, cross-domain recommendations have become a recent research hotspot. Cross-domain recommendations leverage transfer learning techniques to extract rich information from other domains to enhance recommendation performance in sparse domains. Despite extensive research in recent years on cross-domain recommendations and statistical recommendation models, detailed studies applying statistical recommendation models to cross-domain recommendations have not been seen thus far. To fill this gap, this paper proposes a transfer linear model for recommendation systems, aiming to utilize cross-domain recommendation to transfer multi-domain information and leverage the interpretability advantages of linear models to significantly alleviate information silos and cold start problems.

The effectiveness of the proposed algorithm was tested on commonly used datasets in the field of cross-domain recommendations. Experimental results indicate that leveraging source domain information can effectively enhance the model's performance in the target domain, and the proposed algorithm can effectively alleviate cold start and information silo issues in multi-domain scenarios.

Keywords: Recommender System, Transfer Learning, Factorization Machine, Cold Start, Filter Bubble

目 录

第一章 导论

在本章中，将首先介绍选题的背景与研究的意义，该节将先介绍迁移学习和统计模型在推荐系统中的应用，并根据实际问题来说明将迁移学习框架下的统计模型引入推荐系统的优势。然后，从回归分析和推荐系统的角度综述国内外文献，该节将介绍推荐线性回归和跨域推荐的研究历程和现状。最后，介绍研究内容与结构安排。

1.1 选题背景和研究意义

本小节将主要介绍迁移学习在推荐系统中的应用。并根据推荐系统中的经典问题——信息茧房与冷启动，说明将迁移线性模型用于实际推荐场景的优势。

1.1.1 选题背景

（一）跨域推荐——迁移学习在推荐系统中的应用

迁移学习（Transfer Learning）[1]是机器学习和推荐系统方面的一个重要的研究课题。传统的推荐系统通常建立在单一领域的数据上，而实际应用中常常会面临跨领域的推荐问题。

推荐系统中的迁移学习往往称作跨域推荐（Cross-Domain Recommendation）[2]。跨域推荐是当下应对数据稀疏问题较为有效的策略之一。跨域推荐的方法是在源领域（Source Domain）较为丰富的历史数据中获取用户的源领域知识，并通过比如共同用户为中间媒介，来迁移这些用户源领域的信息到目标领域（Target Domain），以此来提升目标领域的推荐准确度。迁移学习可以帮助解决推荐系统中的常见问题，例如冷启动和信息茧房。它可以将源领域的知识和经验应用到目标领域，从而有效地解决这些挑战。

因此，迁移学习在推荐系统中具有重要的意义，可以提供一种有效的方法来改善推荐系统的性能。

（二）推荐系统应用统计模型的必要性——信息茧房与冷启动

在推荐系统中，"信息茧房"和"冷启动"是两个关键的挑战。[3]

1.信息茧房（Filter Bubble）[4]是指在某些推荐系统中，用户往往只接收到与自己兴趣相似的信息，而缺少多样性和新颖性的推荐，从而导致信息的过度过滤。这种现象使得用户在一个狭隘的信息环境中被困住，无法接触到不同观点和意见，而丧失了对更广泛的信息资源的获取能力。信息茧房不仅可能加剧人们的信息偏见和信息孤立，还可能削弱推荐系统对于用户需求的了解和准确性。

统计模型受益于更强的可解释性，易于调整推荐结果，不易陷入信息茧房。而跨域推荐凭借其丰富的多域数据，也能缓解信息茧房。

2.冷启动（Cold Start）[5]是指在推荐系统中，对于新用户或者新物品的推荐困难。因为这些新用户或物品没有足够的行为数据可供推荐系统进行分析和建模。传统的协同过滤方法往往依赖于用户历史行为数据或物品关联信息来做推荐，而新用户和新物品缺乏这些信息，导致推荐系统无法准确地理解其兴趣和特征，从而影响推荐的质量和准确性。

在冷启动瓶颈上，深度推荐模型由于需要大量训练数据往往难以解决冷启动问题。传统统计模型相对深度学习模型对数据的要求较低，如逻辑回归可以利用较少的数据进行建模和预测。进一步的，传统统计模型可解释更强，可以提供对模型预测结果的理解和解释。比如在逻辑回归中，可以通过权重系数来了解每个特征对预测结果的贡献程度。了解模型是基于哪些特征来做出预测对于冷启动问题中的新用户或新物品更为重要。同时，跨域推荐由于能够增强目标域信息，也是缓解冷启动的常见方式。

1.1.2 提出问题

在推荐系统中应用传统统计模型能够缓解推荐系统中的信息茧房和冷启动问题。而迁移学习算法也是解决两个问题的常见做法。但是，到目前为止还没有看到把基于迁移学习的线性模型应用于推荐系统的研究。

1.1.3 研究意义

受到 Li 等人（2022）[6]的启发，本文将迁移学习框架下的回归算法结合常见的推荐模型，并对文章提出的迁移算法做了针对性改变，以适应工业推荐系统，尝试缓解冷启动和信息茧房问题。

本研究致力于研究推荐学习框架下的统计模型在推荐系统场景的应用，这在文献和应用场景仍然未见，但对于解决实际问题是关键的。一方面，迁移学习可以利用更多的信息来提高预测精度并缓解冷启动；另一方面，线性回归模型以其简单的优点能够破除信息茧房并且缓解冷启动。

1.2 国内外文献综述

1.2.1 国外文献综述

（一）统计推荐模型的发展历程

推荐系统是基于统计机器学习的应用，旨在为用户提供个性化的推荐内容。推荐系统中的统计模型最早可以追溯至20世纪90年代。早期的研究工作由Resnick和Varian（1997）[7]提出，他们介绍了推荐系统的基本原理和方法。该研究提出了将最初的协同过滤（Collaborative Filtering，CF）[8]和逻辑回归模型（Logistic Regression，LR）[9]相结合的思想，为后续的研究提供了基础。它基于用户和物品的历史行为数据，利用共现矩阵（Co-Occurrence Matrix）[10]通过挖掘用户之间的相似性以及物品之间的相似性来进行推荐。常见的协同过滤算法包括基于用户的协同过滤和基于物品的协同过滤。

Breese等人（1998）[11]通过实证分析了协同过滤和线性回归模型在推荐系统中的预测算法效果。他们发现，在某些情况下，线性回归模型可以比传统的协同过滤方法更有效。这一研究对于线性回归模型在推荐系统中的应用提供了实证支持和理论指导。

随后，Koren等人（2009）[12]将线性回归模型进一步扩展到矩阵分解模型。他们通过分解用户物品评分矩阵，并利用线性回归模型进行预测，取得了显著的推荐效果。这一研究提供了一种使用线性回归模型进行推荐的新方法，引起了广泛关注。矩阵分解的目的就是通过分解共现矩阵，得到用户和物品的评分，并且通过评分来为用户推荐物品。Agarwal等人（2009）[13]提出的隐语义模型（Latent Factor Model）也是类似的思想。

一年后，Rendle等人（2010）[14]推出了因子分解机（Factorization Machine，FM），这是对线性回归模型的进一步扩展。FM能够更好地捕捉变量之间的交互作

用，提供更准确的预测能力。其基本思想是通过引入隐含因子对特征间的二阶交互进行建模，从而在推荐系统中获得更精细的个性化推荐结果。

在后续的多年内，涌现了大量对 FM 模型的改进工作。在 FM 提出后的一年内，Juan 等人（2016）[15]提出了一种更为强大的模型，称为扩展的因子分解机（Field-aware Factorization Machine，FFM）。FFM 通过引入域特征，考虑特征间的高阶交互。每个特征都与所属字段相关联，使模型能够学习特征在不同字段之间的差异。这种建模方式有效地解决了特征稀疏性和特征交互的问题，进一步提升了推荐算法的性能。FFM 在推荐系统中获得了广泛的应用，取得了显著的提升效果。Pan 等人（2018）[16]提出的 FwFM 是 FM 的扩展，通过为每一对交互域添加权重，显式地建模了不同特征交互的重要性。Pande 等人（2020）[17]提出的 FEFM 则是为每一对交互域创建权重矩阵，进一步提升模型对隐含关系的学习能力。

不像大部分工作专注于对可解释的统计模型进行优化，Cheng 等人（2016）[18]结合统计模型与深度学习，提出了广泛应用的 Wide&Deep 模型，该模型一经提出即在业界取得优秀的结果并且获得了广泛应用。正如其名字一样，模型是由逻辑回归组成的 Wide 部分和深度神经网络组成的 Deep 部分混合而成的模型，其中 Wide 部分主要作用是让模型具有较强的记忆能力，Deep 部分主要是让模型具有较强的泛化能力。这样的结构特点，使模型同时具有了逻辑回归和深度神经网络的优点，能够快速处理并记忆大量历史行为特征，并具有强大的表达能力。

在 Wide&Deep 提出后，涌现了许多融合统计模型与深度学习的模型。Guo 等人（2017）[19]结合 Wide&Deep 模型以及 FM 模型的优点，提出了 DeepFM 模型。它拥有 Wide&Deep 模型的优点，同时借用 FM 模型提取交互作用的思想，使得模型的信息提取能力更进一步。Wang 等人提出的 DCN（2017）[20]和 DCN-V2（2021）[21]等将 Wide&Deep 的 Wide 部分改为了设计的 Cross Network，以显式外积的形式建模特征的交互。xDeepFM（2018）[22]在 Wide & Deep 的基础上添加了设计的交互网络显式构造了有限阶特征组合。

综上所述，统计模型在推荐系统中具有重要的应用价值。通过结合协同过滤、矩阵分解、因子分解机等方法，统计模型能够有效地预测用户对物品的评分或偏好，从而提供个性化的推荐结果。相对深度学习模型，统计模型以其简单的特性

和更强的解释性依旧在应用场景占据半壁江山。对比 FFM、FwFM 等模型是帮助统计模型进行特征的拓展，我们的算法帮助统计模型实现了域的拓展，帮助模型提取弱相关数据域的信息。

（二）迁移学习框架下的线性模型研究现状

与迁移学习在其他领域的广泛应用相比，迁移学习在统计的应用最近才开始被研究。通常的用法是使用迁移学习方法来改进高维线性回归模型的估计性能。[23]

通过利用大量的源数据，Bastani（2021）[24]提出了一种两步迁移学习方法，以提高高维线性回归模型的估计性能。此外，Li 等人（2022）[6]设计了一种基于聚合的算法来检测可迁移的集合。对于高维广义线性模型，Tian 和 Feng（2023）[25]建立了他们估计器的估计和推断结果，还提出了一种数据驱动算法来确定哪个源信息更丰富。与 Tian 和 Feng（2023）利用 Lasso 方法不同，Li 等人（2023）[26]采用了 Dantzig 选择器的思想[27]，并得出了类似的理论结论。

高维多域的情形正是推荐系统实际问题中经常遇到的。且仅需少量样本的统计模型，非常适用于解决推荐系统的冷启动问题。

（三）跨域推荐研究现状

跨域推荐兴起于深度学习时代，而大部分的工作也是基于深度学习模型进行的。较早的跨域推荐工作由 Pan 等人（2010）[28]提出，该方法的主要思想是利用矩阵分解方法在源域进行预训练，接着在目标域对数据进行微调。该研究的出现揭示着迁移学习在推荐系统中的重大作用。

在深度推荐系统发展初期，Man 等人（2017）[29]针对跨领域冷启动场景，提出了 EMCDR 模型，其使用一个显示的映射函数来建模不同领域中用户表征的关系。具体而言，该方法先分别在源领域和目标领域各自训练一个矩阵分解的模型，再将领域共享的用户在源域的表示通过映射函数映射到目标领域。

在此工作之后涌现了大量基于映射的跨域推荐方法。Zhu 等人（2018）[30]提出的 DCDCSR 将映射函数改为由目标领域到标准领域的映射函数，标准领域即通过矩阵稀疏度将源推荐领域和目标推荐领域融合到一起的特征空间，该特征空间被认为是融合了多个领域信息的一般化的标准领域。SSCDR 模型（2019）[31]模型通过设计无监督损失函数，将整个模型转化为半监督学习过程。而 TMCDR 模型

（2021）[32]借鉴了元学习更加细化了整体的学习过程。

除了基于映射的跨域推荐方法外，还有联合训练多个领域共同训练的跨域推荐方法。Singh 等人（2008）[33]将两个领域的数据联合起来进行矩阵分解，通过共享的中间变量实现知识迁移的效果。Jiang 等人（2016）[34]提出的 XPTRANS 改进了之前的矩阵分解方法，对源领域和目标领域两个领域进行联合分解，同时将跨域知识作为约束条件指导矩阵分解，使得联合分解的结果包含一定程度的域间交互知识。

除了矩阵分解方法外，还有大量联合多领域共同训练的方法。Hu 等人（2018）[35]提出了一种新的网络结构称为 CoNet，在两个领域的结构中交互信息，实现领域间对偶知识迁移的效果。Cui 等人（2020）[36]提出的 HeroGraph 对于每个实体在领域内和异构共享图上分别学习两个表示向量，将二者以拼接方式得到了每个领域内用户和物品的表示向量，从而在多个领域上实现更好的推荐。Zhang 等人（2023）[37]提出了一种对偶网络结构，并设计了许多模块，使得模型能够有选择地从源域中挑选对目标域有价值的样本，该模型在实际生产中取得了较好的效果。

本文提出的方法是联合训练多个领域共同训练的跨域推荐方法。对比常见的跨域推荐成果，本文不设计网络结构。本文综合了回归分析和迁移算法提出模型，旨在以可解释的方式实现跨域推荐。

1.2.2 国内文献综述

许等人（2009）[56]全面地总结推荐系统的研究现状，介绍了推荐算法思想、帮助读者了解这个研究领域。吴等人（2022）[57]阐述了推荐系统的几项关键技术，总结了推荐系统的体系结构和性能评价指标，并尝试给出了推荐系统未来研究的重点、难点和热点问题。从二十一世纪十年代互联网兴起后，基于单领域的推荐算法在国内得到了广泛应用。

而随着时代发展以及信息过载现象愈发严重，跨域技术被推向高峰。罗等人（2013）[58]通过矩阵聚类方法来提取矩阵的潜在信息，以实现跨域推荐。陈等人（2017）[59]对跨域推荐进行了介绍，概述跨域推荐算法的相关概念、技术难点，并对跨域推荐的性能分析方法进行详尽的介绍。

但是，未见有文献将统计模型用于跨域推荐。本文将迁移学习框架下的回归模型引入推荐系统，能够极大程度的利用迁移学习和回归模型的优点，解决推荐系统中的顽疾。

1.3 研究内容与方法

本文致力于通过实验证明迁移学习框架下的回归模型能够解决推荐系统中的冷启动和信息茧房问题。

本文首先概述了推荐系统并且详细介绍了推荐系统中常见的两个问题。其次从理论角度介绍迁移学习框架下的回归模型，并给出用于缓解问题的对应算法。最终在推荐系统的热门公开数据上证明回归模型和迁移算法对推荐效果的提升。

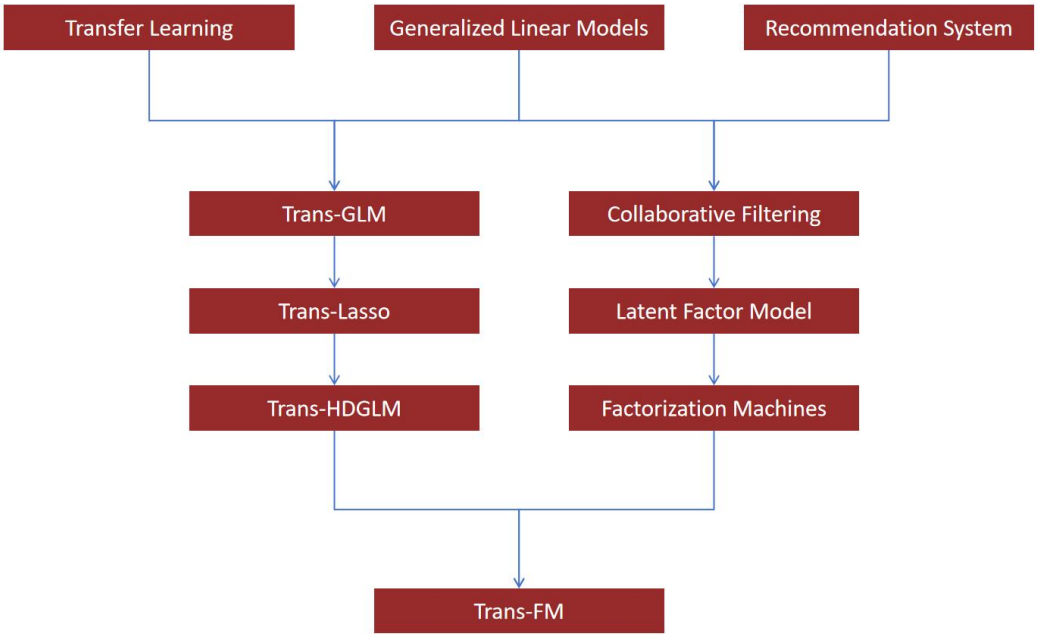

图 1.1 模型构造导图

1.4 创新与不足

1.4.1 本文创新点

本文主要目标在于将迁移学习框架下的回归模型引入推荐系统中，旨在综合迁移学习和回归模型的优点，以极大程度的缓解推荐系统中的信息茧房和冷启动问题。主要创新贡献如下：

1.在以往的工业推荐系统中，可解释的统计模型往往在单个目标域上拟合。本

文将经典统计推荐模型拓展至多领域，在保留可解释性优势的同时，增加了模型的信息提取能力，显著提升了可解释模型的推荐效果。

2.目前跨域推荐的研究中，往往停留在如何将源域信息映射到目标域，或者如何设计网络结构联合训练。本文提出的方法对基础模型没有严格要求，适用于各种推荐线性模型，为跨域推荐提供了新的视角和方法。

1.4.2 研究难点

受限于回归模型的拟合能力，本文提出的算法更适用于需要推荐结果更丰富的场景，而不是极度准确的场景。同时，由于迁移学习的存在，需要多域数据才能实施。

1.5 结构安排

论文的主要讨论由五个章节组成，其余四个章节的具体内容如下：

第二章主要介绍推荐系统及迁移学习的相关知识，并且对推荐系统中常见的冷启动问题和信息茧房问题进行了详细介绍。

在第三章中，介绍了迁移学习框架下线性回归的演化背景，并进一步提出了迁移线性因子分解机模型。该模型旨在缓解冷启动和信息茧房问题。

第四章主要介绍了所做的数值实验，具体分为算法预估性能验证、跨用户冷启动推荐效果验证、跨物品冷启动推荐效果验证和缓解信息茧房效果验证。旨在通过实验证明所提出算法对推荐效果的提升。

结论和未来的工作将在第五章中进行讨论。

第二章 推荐系统概述及常见问题

本章将主要介绍推荐系统、迁移学习和跨域推荐，并给出推荐系统中最常见的两个问题。首先第一节介绍推荐系统及推荐系统中常用的线性模型。第二节介绍迁移学习及常见的迁移学习算法。第三节介绍跨域推荐技术。第四节讲解了推荐系统的两个常见问题，并在后续章节针对这两个问题提出算法。

2.1 推荐系统概述及常用模型

2.1.1 推荐系统概述

推荐系统是一种应用广泛的机器学习技术，用于预测用户可能感兴趣的物品或服务，并向他们提供个性化推荐。推荐系统的主要目标是增强用户体验、提高销售额，并帮助用户发现新的内容。推荐系统通常通过分析用户的历史行为、偏好以及物品的属性来生成推荐。

推荐系统经历了以下几个发展阶段：

1.早期阶段：早期的推荐系统主要依赖于基于内容的推荐和协同过滤推荐。基于内容的推荐利用物品的属性和特征来进行推荐，而协同过滤推荐则是利用用户历史行为数据来发现用户间的相似性，从而进行推荐。

2.初步推广：随着互联网的发展，像 Amazon 和 Netflix 这样的电子商务和娱乐平台开始大规模应用推荐系统。它们采用了更加复杂和有效的推荐算法，如基于协同过滤的推荐、深度学习推荐等，以提升用户体验和销售额。

3.社交网络和移动应用的推广：随着社交网络和移动应用的普及，推荐系统开始更多地考虑用户的社交关系和地理位置等因素。社交网络数据和移动应用数据的加入丰富了推荐系统的特征，使得推荐更加个性化和精准。

4.统计推荐逐渐成熟：随着数据规模的不断增大和机器学习技术的不断进步，个性化推荐逐渐成熟和普及。推荐系统越来越准确地理解用户需求，提供更加个性化的推荐服务，例如音乐应用根据用户的听歌历史推荐新歌曲。

推荐系统在如今的互联网时代扮演着至关重要的角色，不仅可以提高用户体验和满意度，还可以促进商业发展和推动数字内容的传播和推广。

2.1.1 推荐系统常用线性模型

在推荐系统发展的各个阶段，诞生了多种线性推荐模型，下面给出工业系统中常见的推荐线性模型：

1.协同过滤[11]：推荐系统中最流行的技术之一，它分为用户协同过滤和物品协同过滤两种类型。用户协同过滤基于用户行为历史来计算用户之间的相似度，推荐与用户相似的其他用户喜欢的物品；物品协同过滤则基于物品的相似性来向用户推荐相似的物品。协同过滤能够捕捉用户和物品之间的关系，适用于大规模数据集。但是在数据稀疏时难以找到相似用户或物品。

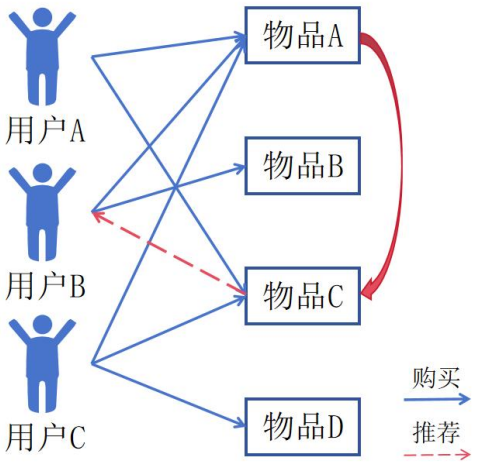

图 2.1 协同过滤算法示意图

2.矩阵分解[12]：一种常见的推荐系统模型，主要用于解决协同过滤中的矩阵稀疏性问题。通过将评分矩阵分解成两个低维稠密矩阵，可以通过乘积重构原始评分矩阵从而进行推荐。矩阵分解能够捕捉隐藏的用户和物品特征、有效解决稀疏性问题，适用于大规模数据集。但是包括对缺失数据处理不够灵活、无法解释潜在特征的含义以及需要调整超参数。

3.线性回归模型[9]：简单而常见的推荐系统模型，通过线性组合特征来进行预测。在推荐系统中，线性回归模型通常会丢失特征之间的高阶关系，难以捕捉数据中的复杂模式和交互作用，因此在推荐系统中的应用受限。线性回归模型简单、解释性强，但是无法处理特征交互性强的情况，预测准确性有限。对于一个有 p 维特征的样本$X = (x_1, \cdots, x_p)$，用于分类的线性回归模型可以表示为

$$P(y = 1|x) = \sigma\left(w_0 + \sum_{i=1}^{p} w_i x_i\right),$$

其中$\sigma(x)=1/(1+e^{-x})$为 Sigmoid 激活函数，$W=(w_0,\cdots,w_p)$为估计参数。

4.二阶多项式模型[38]：一种常见的推荐系统模型，通过引入交叉特征项来考虑特征之间的二阶关系。二阶多项式模型可以有效捕捉特征之间的交互作用，然而由于其高复杂度的缺点，在高维稀疏数据集上往往得不到充分训练。对于一个有 p 维特征的样本$X=(x_1,\cdots,x_p)$，二阶多项式模型可以表示为

$$P(y=1|x)=\sigma\left(w_0+\sum_{i=1}^{p}w_ix_i+\sum_{i=1}^{p}\sum_{j=1}^{p}u_{ij}\,x_ix_j\right),$$

其中$U=\begin{pmatrix}u_{11} & \cdots & u_{1p}\\ \vdots & \ddots & \vdots\\ u_{p1} & \cdots & u_{pp}\end{pmatrix}_{p\times p}$ 为交叉项的估计参数，相对经典线性回归总共有 p^2 个额外参数。

5.因子分解机[14]：一种结合线性模型和矩阵分解的推荐模型。它能够在高维稀疏数据集中学习特征之间的交互关系，通过引入交叉项来捕捉特征之间的二阶关系，从而提高预测性能。因子分解机能够有效处理稀疏数据、防止过拟合，并能够在数据集较小的情况下也能表现良好。对于一个有 p 维特征的样本$X=(x_1,\cdots,x_p)$，二阶因子分解机可以表示为

$$P(y=1|x)=\sigma\left(w_0+\sum_{i=1}^{p}w_ix_i+\sum_{i=1}^{p}\sum_{j=i+1}^{p}v_iv_j\,x_ix_j\right),$$

其中$V=(v_1,\cdots,v_p)$为交叉项的估计参数，相对经典线性回归只有 p 个额外参数。

综上所述，因子分解机在推荐系统中能够兼顾预测性能和效率，相较于二阶多项式模型和线性回归模型具有更好的表现。其高效、高性能和可解释的优点使之成为了最常用的推荐线性模型，后续算法将选择因子分解机作为基础算法并进行改进。

2.2 迁移学习概述及常见算法

2.2.1 迁移学习概述

迁移学习是一种机器学习方法，旨在将从一个领域中学习到的知识迁移到另一个相关但不同的领域中。随着深度学习和大规模数据集的兴起，迁移学习受到越来越多的关注。

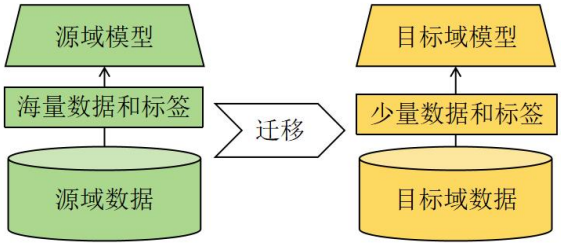

图 2.2 迁移学习示意图

迁移学习随着机器学习和深度学习的进步而快速发展，不断拓展应用领域，在实现知识迁移、解决数据稀缺性等方面发挥了重要作用。

2.2.2 迁移学习常见算法

常见的迁移算法从数据、特征角度，将源域信息迁移到目标域上，以下是几种常见的迁移学习算法：

1.领域自适应算法[39]：旨在解决源领域和目标领域分布不同的问题。该算法通过使源领域和目标领域的数据分布接近，从而提高模型在目标领域上的泛化能力。适用于源领域和目标领域数据特征相似，但分布不同的场景。

2.多任务学习算法[40]：旨在通过同时学习多个相关任务来提升整体性能。在迁移学习中，多任务学习可以通过共享底层特征提取器，使不同任务之间的知识得到迁移。这种方法适用于目标领域拥有少量标注数据或任务之间存在相关性的情况。

3.模型预训练算法[41]：方法是在大规模数据上进行预训练，然后将预训练的模型权重应用于特定任务的微调。模型预训练方法最先由 word2vec 算法提出[41]，这种预训练技术后来被广泛应用于深度学习模型的训练中。常见的模型预训练方法包括自监督学习、无监督预训练、语言模型预训练等。这些方法在处理目标任务数据较少或领域不同时表现出色，能够提供更好的初始化参数和泛化能力。

4.迁移样本选择算法[42]：旨在利用源领域数据中与目标领域相关的样本来进行迁移学习。这种方法通过挑选对目标任务有益的源领域样本，减少噪声和负面迁移，提高模型性能。常见的迁移样本选择方法包括领域匹配采样、关键示例挑选等。该方法适用于源领域中存在大量无关样本或干扰源领域数据的情况，以加速模型收敛和提升准确性。

综上所述，不同的应用场景需要选取不同的迁移学习算法。本文后续章节提出的算法主要是基于迁移样本选择和模型预训练方法实现的。

2.3 跨域推荐

跨域推荐是推荐系统领域中的一个重要分支，目的是在面对多个领域数据时，实现跨领域的信息传递和知识共享，以提升推荐效果。在推荐系统中，根据多个领域信息共享的媒介，可以分为跨用户推荐和跨物品推荐两种[30]，前者代表领域间能共享用户信息，后者代表领域间能共享物品信息。本文后续实验将分别对跨用户场景和跨物品场景进行研究。

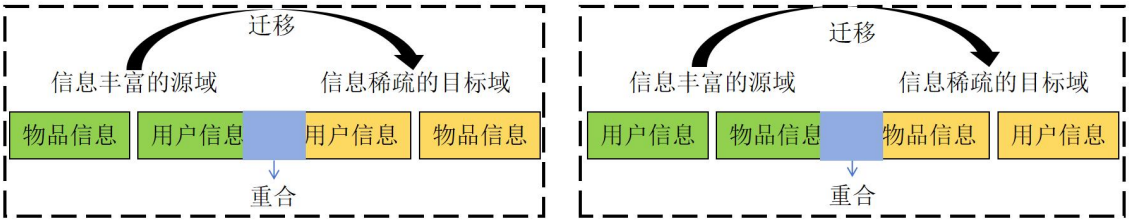

图 2.3 跨用户推荐(左)和跨物品推荐(右)示意图

跨域推荐在近年来迅速发展，经历了以下几个阶段：

1.单域内推荐系统：最初的推荐系统主要关注单一领域内的数据和用户行为，通过分析用户的历史行为数据为其推荐相关内容。

2.多领域推荐系统：随着互联网内容的丰富和推荐系统应用的普及，出现了多领域推荐系统。这些系统尝试将来自不同领域的数据整合在一起，为用户提供更全面的推荐服务。

3.跨域推荐系统：跨域推荐系统超越了多领域推荐系统的范畴，涉及到跨越异构领域的推荐任务。跨域推荐系统通过跨领域的信息共享、迁移学习等技术，将不同领域的知识进行整合，实现更精准的推荐。

4.异构跨域推荐：在实际应用中，存在不同数据类型、不同数据结构的情况，这就需要处理异构数据的跨域推荐系统。异构跨域推荐系统需要考虑数据的语义差异、表示方法差异等问题，提供有效的整合方案。近年来，深度学习技术在跨域推荐中得到广泛应用。深度学习通过学习数据的高级表示，有助于克服跨域数据之间的差异，提高跨域推荐的准确性和效果。

2.4 推荐系统常见问题之冷启动

2.4.1 冷启动

推荐系统的作用是帮助用户发现他们可能感兴趣的物品，提高用户体验和增加平台的粘性。然而，在推荐系统面临一个新用户或新物品时，由于缺乏历史行为数据，系统无法准确为其做出个性化推荐，这就是冷启动问题。

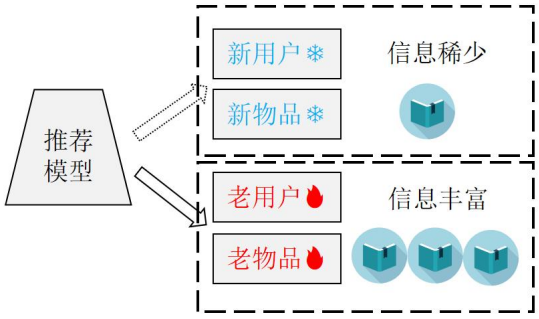

图 2.4 冷启动问题示意图

冷启动主要分为用户冷启动[43]和物品冷启动[44]两种情况。在用户冷启动下，因为没有足够的个人化信息，系统无法根据新用户的行为历史做出个性化推荐，从而影响用户体验和推荐准确度。而物品冷启动则指新物品缺乏足够的历史交互数据，系统无法根据新物品的历史交互数据预测其质量或用户喜好，导致难以推荐给合适的用户群体。

这些问题给推荐系统的效益和性能带来了极大影响。由于缺乏个性化信息会导致推荐的准确度下降，而不能给新用户提供个性化推荐影响用户体验，进而导致用户流失或不满意。

2.4.2 缓解冷启动的方法

为了解决冷启动问题，业界提出了多种解决方案。内容协同过滤[43]可以利用物品的属性信息为新物品建模，减轻物品冷启动问题。流行度推荐[7]则通过向新用户推荐热门物品来缓解用户冷启动问题。主动学习方法[45]可以通过引导用户提供反馈信息，帮助系统更快地了解新用户或新物品，从而缓解冷启动问题。利用社交网络信息也是一个常见的做法，通过分析用户的社交关系来推荐兴趣相似的物品。

跨域推荐方法通过融合不同领域的用户行为数据来缓解冷启动问题。通过迁

移学习技术，建立跨领域的用户兴趣关联，将已有领域的用户行为模式迁移到新领域中，提高对新用户的个性化推荐准确性。此外，利用跨域信息如社交网络数据、用户标签等，辅助推荐系统构建用户画像，降低新用户和物品的冷启动成本。通过跨域数据的共享和分析，推荐系统能够更好地理解用户需求，实现更精准的推荐，从而优化用户体验。本文后续算法将从跨域推荐的角度，缓解冷启动问题。

2.4.3 常用评价指标

对于冷启动任务，评估指标与常见问题的指标相同。常用的评价指标是分类任务的 AUC 指标和回归问题的 RMSE 指标，但是计算范围限定在新物品和新用户上。这些值越高，说明模型对新物品和新用户的推荐效果较好。

在二分类中，绘制一个 ROC 曲线，横轴表示假正例率，纵轴表示真正例率，然后计算曲线下的面积，即可得到 AUC。可以将 AUC 理解为，随机选择一个正负样本对，这个正负样本对正样本预测值大于负样本预测值的概率。当 AUC 越大，表示模型能把更合适的物品推荐给对应的用户。

对于多分类问题，往往使用 *macro*-AUC 指标进行评估。*macro*-AUC 的计算方式略有不同于二分类 AUC。多分类问题中，针对每个类别计算一个分类器的 AUC 值，然后对这些 AUC 值进行平均，即得到 *macro*-AUC 指标。

对于回归模型，使用测试集的均方根误差指标进行评估。数学公式表示为

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum (y\_pred \ - \ y\_true)^2},$$

其中n表示样本数量，y_pred表示预测值，y_true表示真实值。RMSE 的值越小，表示模型的预测误差越小，与真实值更接近。

在后续的实验中，将使用 AUC 和 RMSE 评估算法对冷启动问题的缓和程度。

2.5 推荐系统常见问题之信息茧房

2.5.1 信息茧房

信息茧房问题[46]是指用户接收到的信息过于同质化，缺乏多样性和广度，导致用户只沉浸在特定类型或主题的内容中，无法获得更广泛的信息。这种情况可能导致用户视野狭窄，错失了接触和了解更多新颖事物的机会。

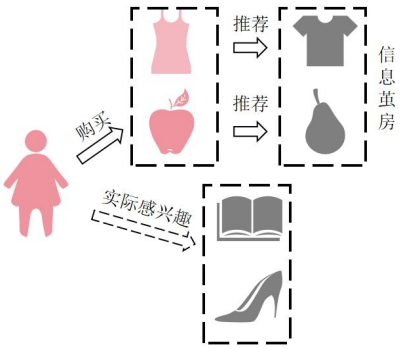

图 2.5 信息茧房问题示意图

信息茧房问题在社交媒体、新闻、音乐等领域尤为常见。例如，在社交媒体平台，如果用户只看到与自己观点一致的内容，即使存在其他不同观点的信息，也无法展现给用户，加剧了信息的同质化。在新闻推荐中，用户长期只浏览某一类型的新闻，可能错过其他领域的重要信息，影响对全面信息的了解。

2.5.2 缓解信息茧房的方法

业界针对信息茧房问题提出了多种解决方案。首先，可以通过引入不同的用户、物品特征来确保推荐结果的多样性，从而打破信息茧房。其次，社交化推荐[47]在推荐过程中考虑用户的社交关系和兴趣，推荐来自不同社交圈子的内容，能提升信息多样性。还有基于深度学习的模型结合内容信息和用户行为，实现精准推荐的同时保持信息的多样性。

跨域推荐通过跨越不同领域和兴趣偏好，为用户推荐更广泛的、多样化的内容，打破信息壁垒，避免陷入信息茧房。通过跨域推荐，用户可以享受到更丰富多彩的信息内容，打破信息狭隘性，拓展信息范围。

2.5.3 常用评价指标

信息茧房一般出现在分类问题中，具体表现为模型预测概率较高的物品间是相似的。

准确的推荐结果代表信息茧房问题并不严重。为了评估推荐结果的召回准确率，我们使用召回率（$Recall@K$）指标。$Recall@K$能评估系统在给用户推荐物品时的召回能力，指的是系统能够在前 K 个推荐结果中恰当地包含用户感兴趣的物品的能力。具体而言，$Recall@K$衡量的是在前 K 个推荐结果中，与用户喜好或历史行为相关的物品数量占总相关物品数量的比例。数学公式表示为

$$Recall@K = \frac{\|PredictSet@K \cap ReferenceSet\|}{\|ReferenceSet\|},$$

其中$PredictSet@K$表示在推荐结果中被认为与用户喜好或历史行为相关的物品的集合。每个用户取前K个概率最大的推荐商品作为其最终的推荐。$ReferenceSet$是指用户在实际情况下真正感兴趣或有过行为互动的物品的集合。这个集合通常是通过用户在测试集中的历史行为数据确定的。

为了评估推荐结果多样性，我们使用汉明距离（Hamming distance）对推荐序列进行评价。数学公式表示为

$$H_{ij}@K = 1 - \frac{Q_{ij}}{K},$$

其中K表示每个用户推荐序列的长度，Q_{ij}为推荐系统给用户 i 和 j 两个推荐列表中相同物品的数量。H_{ij}衡量了不同用户间的推荐结果的差异性，其值越大说明不同用户间的多样性程度更高。在本文后续的实验中，将随机抽取 1000 次用户-用户对并计算汉明距离并取均值记为 *Hamming@K*。

本文后续实验中，将使用 *Recall@K* 和 *Hamming@K* 评估算法对信息茧房问题的缓和程度。

第三章 用于推荐系统的迁移线性模型

本章将介绍用于推荐系统的迁移线性模型，首先介绍迁移线性回归模型，然后针对推荐场景改进该迁移模型。第一节将介绍辅助信息已知情形下，如何让逻辑回归学习辅助信息并提高目标域的预测精度。第二节将进一步介绍在辅助信息混杂的情形下，如何运用算法识别辅助信息。第三节将介绍在推荐系统问题下，如何修正算法提高推荐效果。

3.1 迁移线性模型

在本节中，我们考虑当源域与目标域分布十分接近时的迁移线性回归。

3.1.1 广义线性回归

给定自变量$x \in \mathbb{R}^p$，如果响应变量y遵循广义线性模型（Generalize Linear Model，GLM），则其条件分布的形式为

$$y|x \sim P(y|\alpha) = \rho(y)exp\{yx^Tw - \psi(x^Tw)\},$$

其中$w \in \mathbb{R}^p$为系数，ρ和ψ是一些已知的函数。$\psi'(x^Tw) = E(y|x)$被称为连接函数（inverse link function）[48]。而ψ代表了不同的广义线性模型。例如，在具有高斯噪声的线性模型中，有一个连续的响应变量y和$\psi(u) = u^2/2$；在逻辑回归模型中，响应变量y是离散的独热变量，$\psi(u) = log(1 + e^u)$；在泊松回归模型中，响应变量y是非负的，$\psi(u) = e^u$。

3.1.2 迁移线性模型

在下面的内容中，考虑以下迁移学习问题：假设我们有目标域数据$(X^{(0)}, y^{(0)})$和k个源域数据，第k个源表示为$(X^{(k)}, y^{(k)})$，其中对所有k满足$X^{(k)} \in \mathbb{R}^{n_k \times p}, y^{(k)} \in \mathbb{R}^{n_k}$。$X^{(k)}$的第$i$行和$y^{(k)}$的第$i$个元素分别表示为$x_i^{(k)}$和$y_i^{(k)}$。迁移模型的目标是从源域中获取有用的信息，使得目标域得到更好的模型。假设目标域数据和源域数据都遵循广义线性模型

$$y^{(k)}|x \sim \mathbb{P}(y|x) = \rho(y)\, exp\{yx^Tw - \psi(x^Tw)\},$$

对于源域$k = 0, \ldots, K$，模型具有不同的系数$w^{(k)} \in \mathbb{R}^p$，自变量$x \in \mathbb{R}^p$，以及一些已知的单变量函数ρ和ψ。在本文的讨论中，将目标域系数表示为$\beta = w^{(0)}$。同时，

假设目标域模型是稀疏的，即满足$\|\beta\|_0 = s \ll p$。这意味着只有p个自变量中的s个对因变量有贡献。对于源域，直观地说，如果$w^{(k)}$接近于β，那么第k个源可能对迁移学习有用。

在源域与目标域接近时，所有的源域信息皆可用于目标域的推断。对于$k = 1,\ldots,K$，定义第 k 个域的数据集大小为n_k，具体的算法如下：

算法一: 迁移线性模型[6]

0 输入:

目标域数据$(X^{(0)}, y^{(0)})$, 相关的K个源域数据$\{(X^{(k)}, y^{(k)})\}_{k=1}^{K}$, 正则参数$\lambda_w$、$\lambda_\delta$.

1 迁移步骤:

$$\hat{w}^A \leftarrow \underset{w}{arg\,min}\{\frac{1}{n_A+n_0}\sum_{k=0}^{K}\left[-(y^{(k)})^T X^{(k)} w + \sum_{i=1}^{n_k}\psi(w^T x_i^{(k)})\right] + \lambda_w\|w\|_1\}.$$

2 纠偏步骤:

$$\hat{\delta}^A \leftarrow \underset{\delta}{arg\,min}\{-\frac{1}{n_0}(y^{(0)})^T X^{(0)}(\hat{w}^A + \delta) + \frac{1}{n_0}\sum_{i=1}^{n_0}\psi((\hat{w}^A + \delta)^T x_i^{(0)}) + \lambda_\delta\|\delta\|_1\}.$$

3 输出参数估计:

$$\hat{\beta} \leftarrow \hat{w}^A + \hat{\delta}^A.$$

上述算法首先用利用所有的数据得到参数的初步估计，然后在第二步使用目标数据纠正偏差。注意到迁移样本的分布往往与目标样本的分布不一致，具体体现在它们的协方差矩阵上。若不同，则第一步得估计会有偏差，所以在第二步使用源域数据对估计结果进行纠偏。

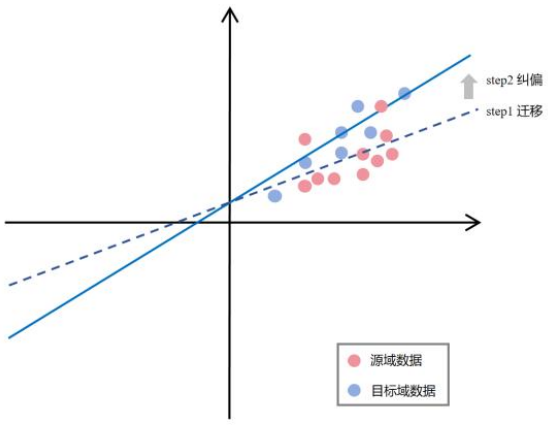

图 3.1 算法一示意图

在本文后续的实验中，将使用交叉熵损失以及 ADAM 算法[49]进行数值优化。

给定自变量 X 和响应变量y，交叉熵损失函数描述如下：

$$L^{bce}(\hat{\beta}; X, y) = -y^T log(\sigma(X\hat{\beta})) - (1-y)^T log(1-\sigma(X\hat{\beta})),$$

其中$\sigma(x) = 1/(1+e^{-x})$为 Sigmoid 激活函数，$\hat{\beta}$为估计的模型参数。

3.2 基于源域识别的迁移线性模型

在本节中，我们考虑当源域混杂时的迁移线性回归，相对第一节，源域中有许多不可用的混杂信息，需要设计算法剔除。选择合适的迁移信息可以避免由于使用不合适的迁移信息而导致的负迁移[50]。

3.2.1 源域信息的识别

为了描述源域是否可以迁移，对于$k = 1, \ldots, K$，定义第k个源域的差异对比度$\delta^{(k)} = \beta - w^{(k)}$，并将其 L1 范数$\left\|\delta^{(k)}\right\|$作为转移水平。我们将$h$水平的迁移集合$\mathbb{A}_h = \{k: \| \delta^{(k)} \|_1 \leq h\}$定义为迁移水平低于$h$的源域。注意，一般来说，水平$h$可以是任何正值，不同的$h$值定义了不同的$\mathbb{A}_h$。实践中，$h$应该相当小，以保证在$\mathbb{A}_h$是有正向效果的。下述提出的算法中使用损失函数的绝对大小来衡量源域的迁移水平。

本节基于多个不同分布源域的假设，这与实际应用问题相近，期望设计算法来对多个分布的源域信息进行识别，获取合适的可迁移信息。本文提出了一种数据驱动的方法来对可迁移域进行识别，如算法二所示。

在算法二中，我们首先将目标域数据随机划分为两个相同大小的子集。不失一般性，这里我们假设n_0是一个偶数，两个子集的大小皆为$n_0/2$。然后，在划分完的目标域上计算参数的简单估计，为后续是否可迁移提供参考。

重要的下一步是，依次将每个源域的数据和子集的数据进行混杂，然后使用算法一得到参数的迁移估计。若该源域有效，则参数的估计应当更准确，但是由于不知道参数的真实情况，需要使用损失来评估估计的精度。并且为了防止泄露，分别在子集的另一部分计算损失。以得到每个源域上的得分。并和目标域上的得分进行比较，若损失小于阈值则认为该源域是可迁移的。

算法二: 源域检测算法

0 输入:

目标域数据$(X^{(0)}, y^{(0)})$, K个源域数据$\{(X^{(k)}, y^{(k)})\}_{k=1}^{K}$, 阈值超参数$\tau$, 正则参数$\left\{\{(\lambda^{(k)[r]})\}_{k=0}^{K}\right\}_{r=1}^{2}$.

1 划分:

将目标域数据$(X^{(0)}, y^{(0)})$随机划分成两个相同大小的子集$\{(X^{(0)[r]}, y^{(0)[r]})\}_{r=1}^{2}$.

2 计算基线参数:

对 *r=1,2* 计算

$$\widehat{\beta}_0^r \leftarrow \underset{\beta}{arg\,min}\{L^{bce}(\beta; X^{(0)[r]}, y^{(0)[r]})\}.$$

3 计算迁移参数:

对$k = 1, \ldots, K$分别使用正则参数$\lambda^{(k)[1]}$和$\lambda^{(k)[2]}$在$\{(X^{(0)[1]}, y^{(0)[1]})\} \cup \{(X^{(k)}, y^{(k)})\}$和$\{(X^{(0)[2]}, y^{(0)[2]})\} \cup \{(X^{(k)}, y^{(k)})\}$上用算法一计算得到参数的估计$\widehat{\beta}_k^1$以及$\widehat{\beta}_k^2$.

4 计算损失:

对于$k = 1, \ldots, K;\ r = 1,2$ 按照交叉熵公式计算

$$\hat{L}^{bce}\left(\widehat{\beta}_k^1; X^{(0)[r]}, y^{(0)[r]}\right), \hat{L}^{bce}\left(\widehat{\beta}_k^2; X^{(0)[r]}, y^{(0)[r]}\right).$$

5 计算得分:

对$k = 0, \ldots, K$计算

$$L_k^{bce} = \frac{\hat{L}^{bce}(\widehat{\beta}_k^1; X^{(0)[2]}, y^{(0)[2]}) + \hat{L}^{bce}(\widehat{\beta}_k^2; X^{(0)[1]}, y^{(0)[1]})}{2}.$$

6 输出可迁移集合:

$$\mathcal{T} = \left\{k: L_k^{bce} \le (1+\tau)L_0^{bce}, k = 1, \ldots, M\right\}.$$

完成了算法二的识别后，将所有的可迁移源域和目标域全部数据组合，再将数据放入算法一，即可得到最终参数估计。整体的迁移逻辑回归算法如算法三所示。

算法三能够帮助单域模型利用其他源域的信息，实现跨域推荐。通过跨领域的推荐，能够拓展推荐系统的边界与深度，缓解冷启动和信息茧房问题。但是，由于推荐任务对模型预测准确性有较高的要求，而经典的线性模型由于表达能力

无法完全胜任，所以算法需要针对推荐系统做出改变。

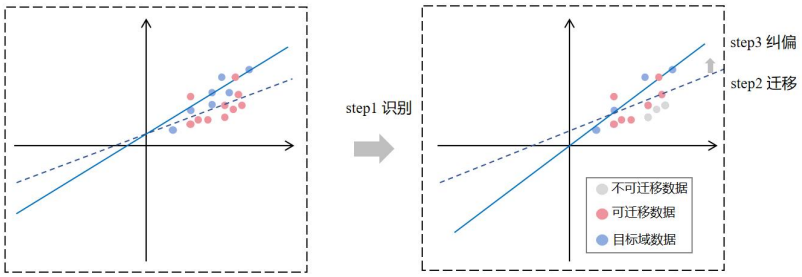

图 3.2 算法三示意图

算法三: 迁移线性回归

1 输入:

目标域数据$(X^{(0)}, y^{(0)})$, 相关的K个源域数据$\{(X^{(k)}, y^{(k)})\}_{k=1}^{K}$, 阈值超参数$\tau$, 正则参数$\left\{\{(\lambda^{(k)[r]})\}_{k=0}^{K}\right\}_{r=1}^{2}$、$\lambda_w$、$\lambda_\delta$.

2 获取迁移集合:

代入所需数据及参数计算算法二，得到可迁移集合$\hat{\mathcal{T}}$，集合中的源域数据$\{(X^{(k)}, y^{(k)}), k \in \hat{\mathcal{T}}\}$认为是可迁移的.

3 数据划分:

将目标域数据和迁移源域数据按照数据域记为两个集合

$$\mathcal{A} = \{(X^{(0)}, y^{(0)})\}, \mathcal{B} = \{(X^{(k)}, y^{(k)}), k \in \hat{\mathcal{T}}\}.$$

4 迁移:

$$\hat{w}^A \leftarrow \underset{w}{arg\,min}\{\frac{1}{\|\mathcal{A}\|+\|\mathcal{B}\|}\sum_{(X^{(k)},y^{(k)})\in\mathcal{A}\cup\mathcal{B}} L^{bce}(w; X^{(k)}, y^{(k)}) + \lambda_w\|w\|_1\}.$$

5 纠偏:

$$\hat{\delta}^A \leftarrow \underset{\delta}{arg\,min}\{\frac{1}{\|\mathcal{A}\|} L^{bce}(\hat{w}^A + \delta; X^{(0)}, y^{(0)}) + \lambda_\delta\|\delta\|_1\}.$$

6 输出参数估计:

$$\hat{\beta} \leftarrow \hat{w}^B + \hat{\delta}^B.$$

3.3 用于推荐系统的迁移线性模型

在推荐任务中，更加关心模型预测结果的准确性。在此情形下，经典的做法是使用因子分解机对线性回归模型进一步扩展。因子分解机能够更好地捕捉变量

之间的交互作用，提供更准确的预测能力。其基本思想是通过引入隐含因子对特征间的二阶交互进行建模，从而在推荐系统中获得更精细的个性化推荐结果。

与传统的逻辑回归模型不同，因子分解机考虑了特征的交集，并对所有特征变量的相互作用进行建模，这使得因子分解机有极强的预估性能。此外，因子分解机模型还具有能够在线性时间内进行计算、集成多种类型的信息以及能够与许多其他模型集成的优点。

对于一个样本$X=(x_1,\cdots,x_p)$，阶数为k的因子分解机可以表示为

$$P(y=1|x)=\sigma\left(w_0+\sum_{i=1}^{p}w_ix_i+\sum_{i=1}^{p}\sum_{j=i+1}^{p}\langle v_i,v_j\rangle x_ix_j\right),$$

该因子分解机相对逻辑回归模型额外多了估计参数$V=(v_1,\cdots,v_p)$，而$\langle\cdot,\cdot\rangle$是两个大小为$k$的向量的点积

$$\langle v_i,v_j\rangle=\sum_{f=1}^{k}v_{i,f}\cdot v_{j,f},$$

V中的第i行v_i描述了具有k个因子的第i个变量。k是一个定义因子分解维数的超参数。直观来说，因子分解机就是为每个特征分配了一个维数为k的嵌入表征，当该特征与其他特征交叉时，用该嵌入作为该特征的表征。该维数越高代表交互的信息越多，在后续的实验中可以看到，对于不同的问题，k的值越大不一定会使得模型表现更优秀。

相对逻辑回归模型，因子分解机实现了自动特征衍生。因子分解机使用$\langle v_i,v_j\rangle$模拟第 i 个变量和第 j 个变量之间的相互作用。在参数估计上，因子分解机不是对每个交互使用自己的模型参数，而是通过分解它来建模交互，这是允许在稀疏性下对高阶交互作用的准确参数估计的关键点。在表达能力上，对于任何正定矩阵W，都存在一个假设k足够大的矩阵V，满足 $W=V\cdot V_t$。这表明，如果k选择足够大，因子分解机可以表示任何相互作用矩阵W。然而，在稀疏设置下，通常应该选择一个小的k，因为没有足够的数据来估计复杂的交互W。限制k的大小能使因子分解机的表达导致更好的泛化，从而改进稀疏性下的交互矩阵。[14]

因子分解机自从 2010 年被提出后，由于易于合并交叉特征、可以处理高维稀疏数据，并且效果不错，在推荐系统及广告点击率预估等领域得到了大规模的使用。将因子分解机引入迁移线性回归模型框架中，有利于大幅提升模型的预测效

果。

结合算法三以及因子分解机算法，本文提出用于推荐系统的迁移因子分解机具体为算法四。

算法四: 迁移因子分解机

1 输入:

目标域数据$(X^{(0)}, y^{(0)})$, 相关的 K 个源域数据$\{(X^{(k)}, y^{(k)})\}_{k=1}^{K}$.

2 获取迁移集合:

代入所需数据及参数使用逻辑回归模型计算算法二, 得到可迁移集合$\hat{\mathcal{T}}$, 集合中的源域数据$\{(X^{(k)}, y^{(k)}), k \in \hat{\mathcal{T}}\}$认为是可迁移的.

3 数据划分:

将目标域数据和迁移源域数据按照数据域记为两个集合

$$\mathcal{A} = \{(X^{(0)}, y^{(0)})\}, \mathcal{B} = \{(X^{(k)}, y^{(k)}), k \in \hat{\mathcal{T}}\}.$$

4 迁移:

$$\hat{y}(w, u, X) \leftarrow w_0 + \sum_{i=1}^{n} w_i x_i + \sum_{i=1}^{n} \sum_{j=i+1}^{n} \langle u_i, u_j \rangle x_i x_j,$$

$$\hat{w}^C, \hat{u}^C \leftarrow \underset{w,u}{arg\,min} \{ \frac{1}{\|\mathcal{A}\| + \|\mathcal{B}\|} \sum_{(X,y) \in \mathcal{A} \cup \mathcal{B}} L^{bce}(\sigma\left(\hat{y}(w, u, X)\right), y) \}.$$

5 纠偏:

$$\hat{y}(\delta, \theta, X) \leftarrow w_0 + \delta_0 + \sum_{i=1}^{n} (w_i + \delta_i) x_i + \sum_{i=1}^{n} \sum_{j=i+1}^{n} \langle v_i + \theta_i, v_j + \theta_j \rangle x_i x_j,$$

$$\hat{\delta}^C, \hat{\theta}^C \leftarrow \underset{\delta,\theta}{arg\,min} \{ \frac{1}{\|\mathcal{A}\| + \|\mathcal{B}\|} \sum_{(X,y) \in \mathcal{A} \cup \mathcal{B}} L^{bce}(\sigma\left(\hat{y}(\delta, \theta, X)\right), y) \}.$$

6 输出参数估计:

$$\hat{\beta} \leftarrow \hat{w}^C + \hat{\delta}^C, \hat{v} \leftarrow \hat{u}^C + \hat{\theta}^C.$$

在选择迁移样本的步骤中，考虑到经典的算法即可得到较好的迁移域选择效果，所以依旧使用逻辑回归进行选择。在后续的估计步骤中，先混杂源域数据和目标域数据，估计得到因子分解机参数的初步估计。最终进行纠偏，并得到所有参数的估计。

第四章 实验分析

在本章中，先在模拟数据集上验证了迁移因子分解机的预估能力。然后在三个常见的跨域推荐真实数据集上进行了大量实验，从多个角度评估所提出的用于推荐系统的迁移线性回归模型的实践性能：在 MovieLens-Netflix 数据集上，我们研究了模型跨物品的冷启动能力；在 Douban 数据集上，我们研究了模型跨用户的冷启动能力；在 Amazon 数据集上，我们研究了模型的召回效果，并评估了信息茧房指标。

4.1 模拟数据验证

在本节中，我们设计数值实验证明所提出的用于推荐系统的迁移线性模型在常见设置下比经典迁移线性模型更优。

4.1.1 模拟数据设置

我们设置总共有 $K = 10$ 个不同的源域。其中，有两种类型的源域分别记为$\mathbb{A}_h$和$\mathbb{A}_h^c$。来自$\mathbb{A}_h$的域与目标域相似，而来自$\mathbb{A}_h^c$与目标域有很大不同，且生成的样本都是未知来源的。每个源域设置样本量$n_k = 100$ 和目标域样本量$n_0 = 200$。

目标域数据和源域数据的维度设置为$p = 2000$。目标域基础参数设置为$\beta = w^{(0)} = \left(0.5 \cdot 1_s, 0_{p-s}\right)^T$, s 为 20。对于任何可迁移的源域数据$k \in \mathbb{A}_h$，将基础参数设置为$w^{(k)} = \beta + (h/p)\mathcal{R}_p^{(k)}$。对于不适合迁移的数据，将参数$w^{(k)}$的第 j 个分量设为

$$w_i^{(k)} = \begin{cases} 0.5 + 2h\mathcal{R}_i^{(k)}/p, & i \in \{s+1, \ldots, 2s\} \cup S^{(k)}, \\ 2h\mathcal{R}_i^{(k)}/p, & otherwise, \end{cases}$$

其中$S^{(k)}$是从$\{2s+1, \ldots, p\}$中随机生成大小为s的集合，$\mathcal{R}_{\mathrm{i}}^{(\mathrm{k})}$表示为$p$维的独立同分布噪声变量（每一维等概率的设置为−1 或 1）。

为了模拟变量间的交互作用，设置交互参数v。在目标域上，交互参数设置为$v^{(0)} = \left(0.5 \cdot 1_t, 0_{p-t}\right)^T$, t 代表交互作用的大小。对于任何可迁移的源域数据$k \in \mathbb{A}_h$，参数设置为$v^{(k)} = \beta + (h/p)\mathcal{R}_p^{(k)}$。对于不适合迁移的数据，将参数$v^{(k)}$的第 i 个分量设为

$$v_i^{(k)} = \begin{cases} 0.5 + 2h\mathcal{R}_i^{(k)}/p, & i \in \{s+1, \ldots, 2s\} \cup V^{(k)}, \\ 2h\mathcal{R}_i^{(k)}/p, & otherwise, \end{cases}$$

其中$V^{(k)}$是从$\{2s+1, \ldots, p\}$中随机生成大小为s的集合，$\mathcal{R}_i^{(k)}$是噪声变量。

目标域自变量皆满足$x^{(0)} \overset{i.i.d.}{\sim} N(0_p, \Sigma)$，其中$\Sigma = [\Sigma_{ij}]_{p\times p}$并且$\Sigma_{ij} = 0.9^{|i-j|}$。对于源域自变量，取自$x^{(k)} \overset{i.i.d.}{\sim} N(0_p, \Sigma + \varepsilon\varepsilon^T)$，其中$\varepsilon \sim N(0_p, 0.3^2 I_p)$。在源域及目标域上，因变量皆由下述线性模型生成

$$y = \sum_{i=1}^{p} w_i x_i + \sum_{i=1}^{p} \sum_{j=i+1}^{p} v_i v_j x_i x_j + \varepsilon,$$

其中w_i，v_i和x_i分别为参数w，参数v和自变量x的第 i 个分量，$\varepsilon \overset{i.i.d.}{\sim} N(0,1)$。

我们在不同的 h、K、t 下训练，然后计算β的 L2 估计误差，并且计算模型预估结果$\hat{y}$和真实值y的 AUC。所有实验均重复 200 次。

4.1.2 数值实验结果

在下图，我们提供了分类任务在不同设置下 200 次实验的平均结果。

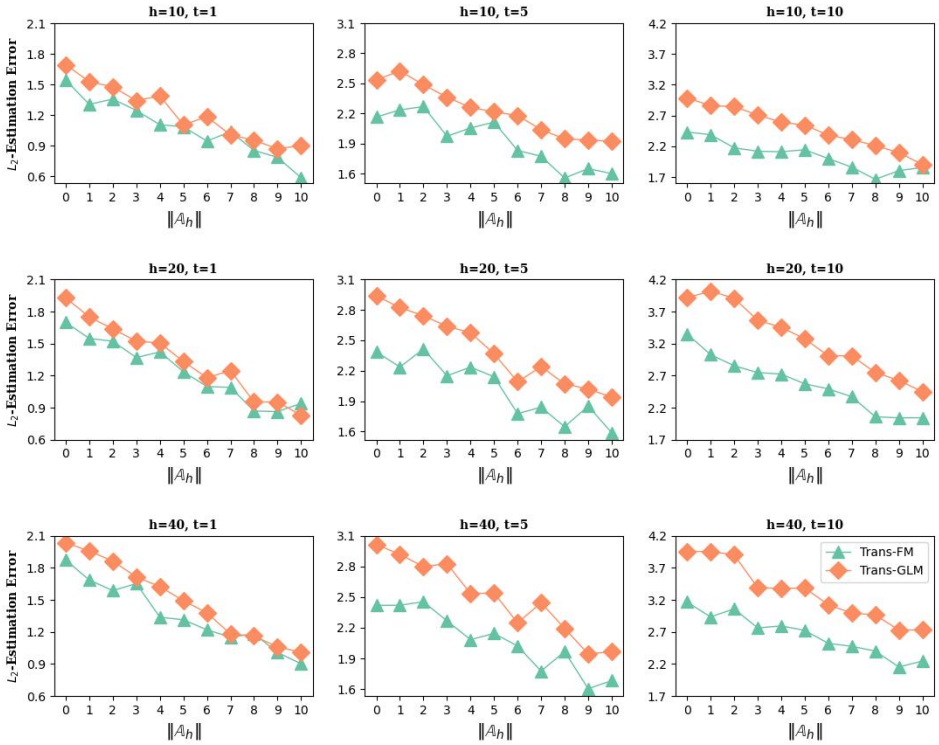

图 4.1 不同设置下估计误差图

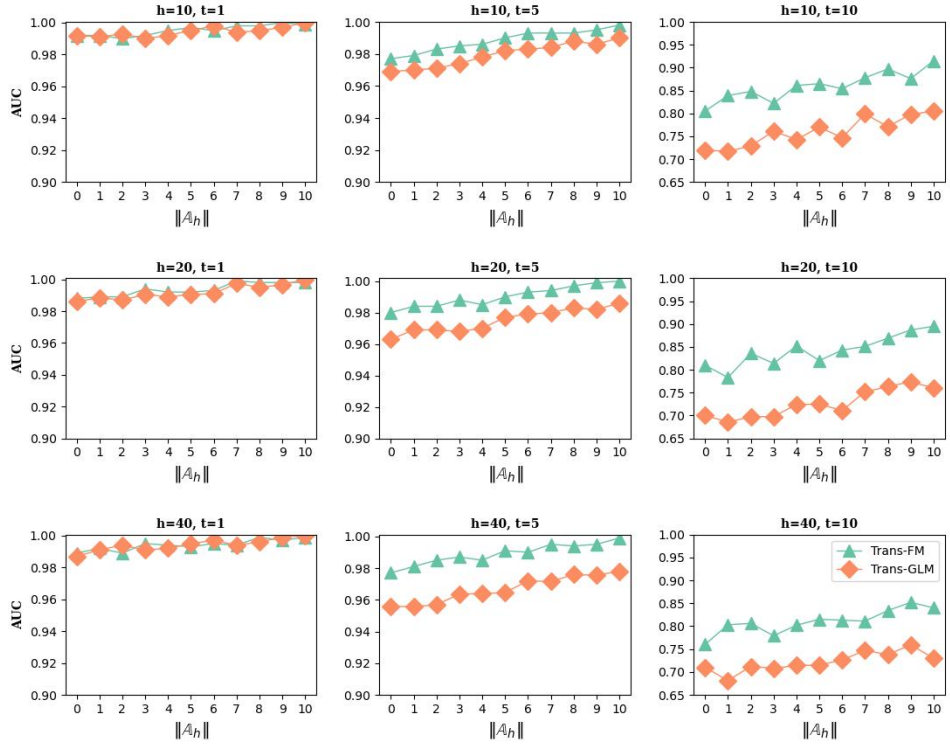

图 4.2 不同设置下 AUC 值图

其中只评估了非交叉参数$w^{(k)}$上的估计误差，Trans-GLM 代表经典迁移线性回归模型，Trans-FM 则代表所提出的迁移因子分解机模型。从图中我们可以得出结论：

1.在有交叉信息的情况下，Trans-FM 的效果优于 Trans-GLM。

2.随着特征交叉程度 t 的增加，Trans-FM 的效果越来越好。

3.在有较强交叉信息的条件下，Trans-GLM 会难以估计真实参数。

4.随着交叉程度的增加，虽然经典线性回归模型能够达到相近的参数估计效果，但是估计效果会逐渐差于有交叉估计的 FM 模型。在十分复杂的特征交叉情况下，Trans-FM 的预测效果远好于 Trans-GLM。

4.2 跨物品冷启动实验

4.2.1 数据集介绍

为了验证模型的跨物品的冷启动效果。我们使用 MovieLens-Netflix[29]数据集，包含两个平台不同的用户和完全相同的电影数据，Netflix 平台和 MovieLens 平台分享了超过 5000 部电影的评分。这两个平台上的用户差异很大，形成了一个共享物品的跨域推荐场景。MovieLens 数据集可以从 https://grouplens.org/datasets/movielens/ 中获取。Netflix 数据集可以从

https://www.kaggle.com/netflix-inc/netflix-prize-data 获取。

MovieLens 数据集是 GroupLens Research 从 MovieLens 网站收集的。MovieLens 是一个经典的电影推荐数据集，经常被用来进行推荐系统和机器学习的实验和研究。它包含了用户对电影的评分和电影的元数据信息。根据不同的收集时间，该数据集的大小不同。我们使用了 2003 年 2 月发布的 MovieLens 1M Dataset。整个数据集由三部分组成：评分数据、用户数据、电影数据。其具体介绍如下：

1.MovieLens 评分数据：包括用户 ID、电影 ID、评分、时间戳。同时，每个用户至少有 20 条评分，且可以对一部电影重复评分。

表 4.1 MovieLens 评分数据描述表

	用户 ID	电影 ID	评分	时间戳
描述	用户编码	用户编码	仅整数星级	以秒为单位
最小值	1.0	1.0	1.0	-
中位数	3070.0	1835.0	4.0	-
最大数	6040.0	3952.0	5.0	-
均值	3024.5	1865.5	3.5	-
标准差	1728.4	1096.0	1.1	-
举例 1	3841	3200	5	965996234
举例 2	6036	30	4	956712526

2.MovieLens 用户数据：所有人口统计信息均由用户自愿提供。所有用户皆有人口统计信息。主要包括用户 ID、性别、年龄、工作类型。

表 4.2 MovieLens 用户数据描述表

	用户 ID	性别	年龄	工作类型
描述	用户编码	用户的性别	用户的年龄	工作类型编码
最小值	1.0	-	1.0	0.0
中位数	3020.5	-	25.0	7.0
最大数	6040.0	-	56.0	20.0
均值	3020.5	-	30.6	8.1
标准差	1743.7	-	12.9	6.3
举例 1	3203	M	35	0
举例 2	4804	F	35	1

3.MovieLens 电影数据：包括电影 ID、标题、类型。

表 4.3 MovieLens 电影数据描述表

	电影 ID	标题	类型
描述	用户编码	电影标题	每个电影可以有多个类型
计数	3883	3883	3883
最小值	1.0	-	-
中位数	2010.0	-	-
最大数	3952.0	-	-
均值	1986.0	-	-
标准差	1146.7	-	-
举例 1	2490	Payback (1999)	Action|Thriller
举例 2	293	Leon(1994)	Crime|Drama|Romance

Netflix 数据集是一个广泛使用的用于电影推荐和数据分析的公开数据集，它包含了 Netflix 平台上的真实用户观影行为数据，该行为数据是以时间戳和打分的形式给出的。这个数据集最初由 Netflix 举办的一个竞赛活动 Netflix Prize 而创建，该活动旨在通过算法改进电影推荐系统。整个数据集由两部分组成，评分数据、电影数据。其具体介绍如下：

1.Netflix 评分数据：包括用户 ID、电影 ID、评分、时间。

表 4.4 Netflix 评分数据描述表

	用户 ID	电影 ID	评分	时间
描述	用户编码	电影编码	仅整数星级	日期时间
计数	100480507	100480507	100480507	100480507
最小值	6.0	1.0	1.0	1999-11-11
中位数	1319012.0	9051.0	4.0	-
最大数	2649429.0	17770.0	5.0	2005-12-31
均值	1322488.5	9070.9	3.6	-
标准差	764536.8	5131.9	1.1	-
举例 1	11696	1600698	3	2001-07-17
举例 2	468	2552314	3	2004-05-12

2.Netflix 电影数据：包括电影 ID、标题、年份。

表 4.5 Netflix 电影数据描述表

	电影 ID	标题	年份
描述	用户编码	电影标题	电影发布年份
计数	17434	17434	17434
最小值	1.0	-	-
中位数	8873.0	-	-
最大值	17770.0	-	-
均值	8879.0	-	-
标准差	5129.5	-	-
举例 1	8514	Fourplay	20011
举例 2	12016	Classic Cartoon Favorites	2005

在后续实验中，以 MovieLens 为源域，以 Netflix 为共享物品的目标域。对比源域和目标域的部分数据可以发现，在整体评分的分布上源域和目标域十分接近，但是源域的数据量远远大于目标域，而且源域数据的产生时间更晚一些。相近的评分代表迁移的可行性，产生时间上的不一致这对迁移算法提出了一定的挑战。

4.2.2 数据预处理

使用的 MovieLens 数据集和 Netflix 数据集分别来自不同的评级网站。数据集的电影 ID 使用了不同的编码方式，故需要使用共同的电影标题来进行对齐，将没有在 MovieLens 和 Netflix 同时出现的电影删除，只保留共同出现的电影。最终保留的数据统计如下：

表 4.6 MovieLens-Netflix 统计对比表

统计	MovieLens 数据集	Netflix 数据集
电影数量	2,047	2,047
用户数量	6,040	463,577
评分数量	628,020	32,513,049

与深度学习其他领域如计算机视觉、自然语言处理等不同，推荐系统场景下，特征工程仍然对模型效果起着至关重要的作用。数据决定了效果的上限，算法只能决定逼近上限的程度，而特征工程则是数据与算法之间的桥梁。

所使用的基础特征包括一下几个角度：

1.用户特征：交互过的物品数，平均评分，评分标准差等。

2.物品特征：交互过的用户数，平均评分，评分标准差，类别平均评分，发布年份等。

3.用户-物品交互特征：用户对物品所属类别的交互数，平均评分（类别交互平均评分），评分标准差等。

4.高阶特征：用户与物品所属类别的交互次数在该用户交互中的比例。

为了便于可解释性分析并利于优化，所有特征皆使用 z-标准化，即对于输入特征 z，输出标准化后的特征$(z - z_{mean})/z_{std}$。

得到的核心特征的统计描述如下：

表 4.7 目标域 MovieLens 特征分布表

特征域	用户域			物品域					交互域	
特征	评分均值	评分标准差	评分计数	评分均值	评分标准差	评分计数	类别平均分	年份	交互平均分	交互次数
最小值	1.0	0.00	7	1.0	0.00	1	1.32	1919	1	1
中位数	3.5	1.00	226	3.6	0.97	693	3.55	-	3.6	7
最大值	5.0	1.91	1487	4.8	2.12	5295	4.33	2005	5.0	303
均值	3.5	1.01	287	3.5	0.98	941	3.56	-	3.5	18.27
标准差	0.4	0.19	226	0.5	0.11	891	0.32	-	0.7	26.83
举例 1	3.8	0.80	82	3.8	0.866	1627	3.553	1987	4.0	1.0
举例 2	3.6	1.30	28	3.8	1.037	1751	3.818	1995	-	-
举例 3	4.2	0.77	235	4.0	0.817	107	3.795	1941	4.3	87.0

表 4.8 源域 Netflix 特征分布表

特征域	用户域			物品域					交互域	
特征	评分均值	评分标准差	评分计数	评分均值	评分标准差	评分计数	类别平均分	年份	交互平均分	交互次数
最小值	1.0	0.00	1	1.7	0.71	60	2.39	1919	1.00	1
中位数	3.6	0.92	181	3.6	0.99	47504	3.62	-	3.67	8
最大值	5.0	2.83	2237	4.5	1.48	297370	4.30	2005	5.00	418
均值	3.6	0.93	231	3.6	0.99	62285	3.64	-	3.65	18
标准差	0.4	0.21	199	0.3	0.09	53626	0.19	-	0.64	25
举例 1	3.9	0.88	110	3.7	1.09	79325	3.58	2000	3.769	25.0
举例 2	3.3	0.73	370	3.8	0.93	60861	3.74	1993	3.375	31.0
举例 3	3.3	0.85	206	3.7	1.11	141442	3.79	1987	-	-

其中有缺失的交互数据将会使用 0 填充。从表中可以发现，目标域特征与源域特征的整体分布接近但是在一些特征（如电影评论数）上有显著差异，而源域的数据量更多。为了确认引入源域能提高目标域的模型效果，需要对迁移的可行性进一步分析。

为了评估源域与目标域的差距，我们需要找到重要的特征，并且对比其在源域和目标域上的区别。我们先在目标域上使用 CatBoost[51]模型进行多分类建模，然后对模型使用 SHAP 分析[52]。SHAP（Shapley Additive Explanations）分析是一种解释机器学习模型预测结果的方法，它提供了对特征重要性的解释，帮助我们理解模型对不同特征的依赖程度。SHAP 分析基于合作博弈论中的 Shapley 值概念而来，这个概念用于衡量参与合作的成员对某个合作成果的贡献。SHAP 分析的核心思想是对每个特征进行排列组合，并计算每种排列组合下特征的贡献值。这些

贡献值的平均值就是特征的 SHAP 值，可以用来解释特征对预测结果的影响。

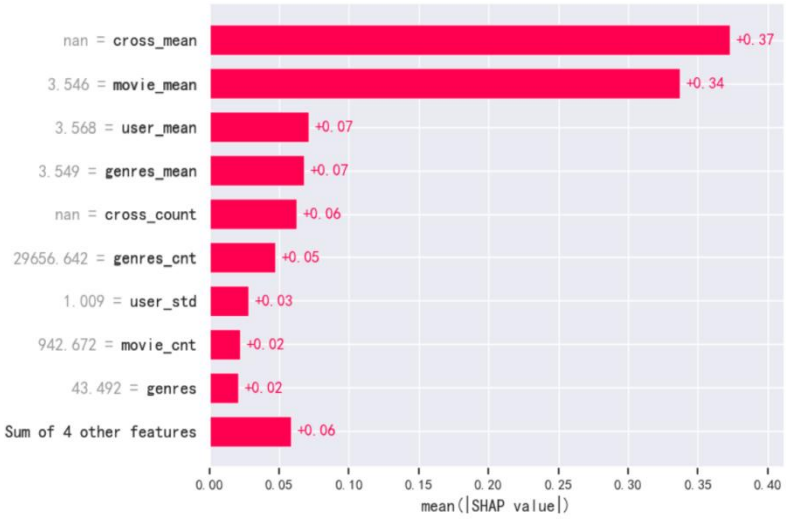

图 4.3 SHAP 分析结果

其中，特征的 SHAP 值越高代表其对预测结果更重要。重要性最高的是类别交互平均评分（cross_mean）、电影平均评分（movie_mean）、用户平均评分（user_mean）和类别平均评分（genres_mean）。

之后，在重要特征上对比目标域数据和源域数据的 KDEPlot（Kernel Density Estimation Plot）。KDEPlot 用于估计和展示连续变量的概率密度函数。它通过核密度估计方法来估计数据的概率密度，并以平滑的曲线形式呈现。KDEPlot 常用于分析单个变量的分布情况，其主要目的是揭示数据的概率密度分布，包括峰值、形状和变化趋势。

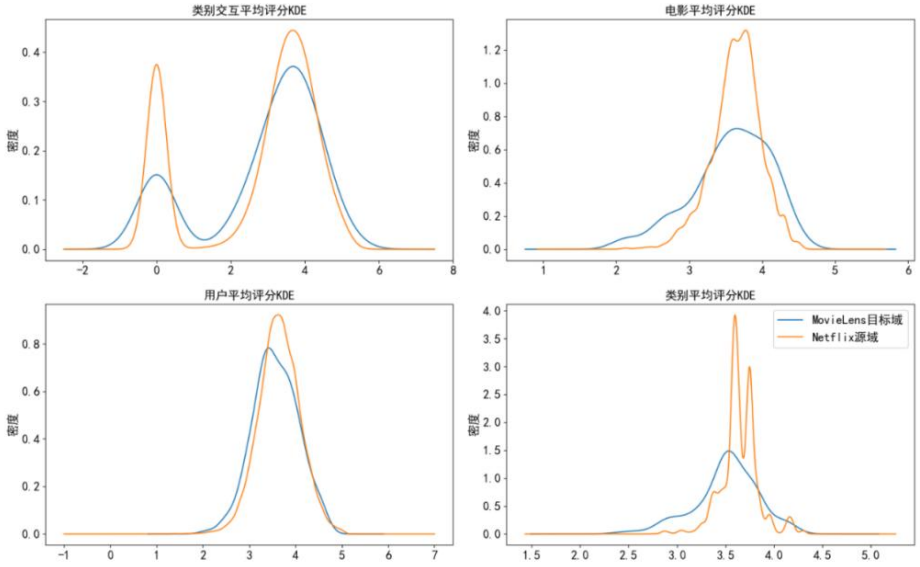

图 4.4 目标域及源域 KDE 分布对比

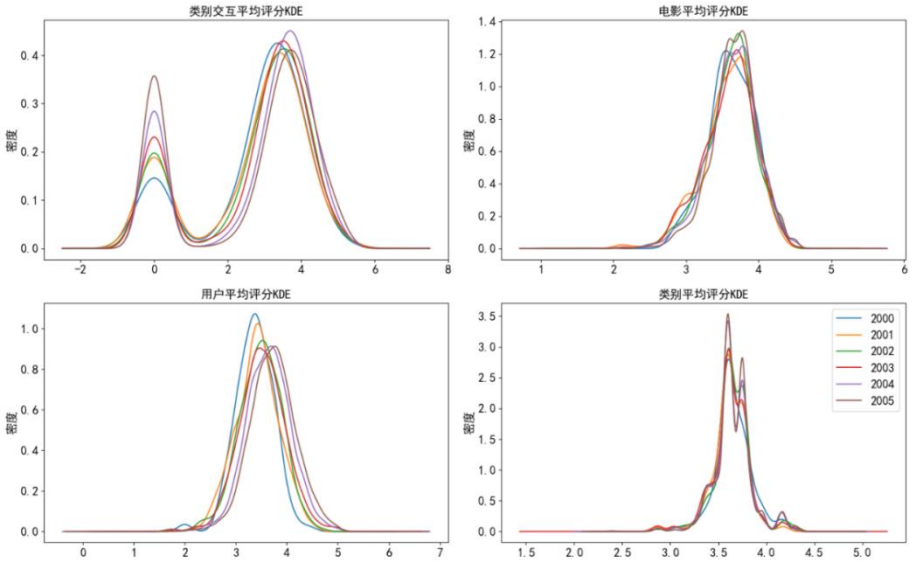

图 4.5 源域间 KDE 分布对比

从图 4.4 中可以发现，在部分重要特征上（如类别交互平均评分、用户平均评分）上，目标域和源域的分布十分接近，而在另外一部分特征上，目标域和源域则相差甚远。这说明源域中有一些信息适合迁移，而一些信息需要进行剔除。在这种情况下，选择合适的源域并在迁移后进行纠偏尤为重要。

在执行源域选择算法时，还需要将全部的源域分为多个来源不同的子源域。考虑到一部电影的评分分布会随着时间的推移产生变化，并且目标域和源域评分产生的年份并不重合，我们将按评分产生的年份将 Neflix 源域数据分为多个子集，并从中选择合适的源域进行迁移。

对比源域上不同年份的评分分布，从图 4.5 中可以发现，随着时间的推移，整体评分逐步轻微提升。而由于源域和目标域是在不同的年代产生的评分数据，对分布差异较大的源域数据进行剔除。在本节最后会展示选择得到的源域的时间分布。

4.2.3 实验结果

在 MovieLens-Netflix 数据集上，为了评估所提框架在跨域推荐任务上的有效性和效率，我们以一定比例随机删除了目标域中一小部分实体的所有打分信息，并将它们作为测试集进行推荐。为了实验的严格性，我们对测试集设置了不同的占比，即 10%、20%、30%、40%和 50%。各种测试集电影占比设置下的 10 次 5 折交叉验证实验平均结果如下：

表 4.9 跨物品冷启动回归 RMSE 结果表

		10%	20%	30%	40%	50%
	AVE	1.0431	1.0321	1.0279	1.0319	1.0292
	GLM	0.9523	0.9619	0.9627	0.9694	0.9743
	Trans-GLM	0.9214	0.9232	0.9252	0.9223	0.9325
K=1	FM	0.9332	0.9354	0.9417	0.9453	0.9476
K=1	Trans-FM	0.9027	0.9059	0.9045	0.9071	0.9104
K=2	FM	0.9343	0.9354	0.9429	0.9448	0.9502
K=2	Trans-FM	0.9009	0.9045	0.9028	0.9066	0.9090
K=3	FM	0.9300	0.9366	0.9402	0.9445	0.9451
K=3	Trans-FM	0.8970	0.8997	0.9001	0.9034	0.9050

表 4.10 跨物品冷启动多分类 macro-AUC 结果表

		10%	20%	30%	40%	50%
	GLM	0.6061	0.6045	0.6040	0.6033	0.6019
	Trans-GLM	0.7024	0.7004	0.6991	0.6982	0.6968
K=1	FM	0.6121	0.6113	0.6100	0.6119	0.6092
K=1	Trans-FM	0.7222	0.7211	0.7215	0.7201	0.7203
K=2	FM	0.6119	0.6115	0.6110	0.6113	0.6101
K=2	Trans-FM	0.7232	0.7220	0.7213	0.7220	0.7210
K=3	FM	0.6122	0.6119	0.6115	0.6113	0.6109
K=3	Trans-FM	0.7240	0.7230	0.7235	0.7225	0.7223

其中 AVE 代表用源域内的平均评分来作为目标域的预测结果，该模型不考虑特征，没有个性化的结果。GLM 为经典线性模型，只使用目标域的数据进行训练。FM 相对 GLM 拓展了交叉项，K 代表 FM 交叉项系数的维数，越大代表越注重交叉项的表征。Trans-X 代表将源域数据迁移到目标域，并在目标域评估。

从这两个表中，我们可以看到，AVE 推荐方法作为一种简单的方法表现出最差的性能。所提出的 Trans-FM 模型在 RMSE 和 AUC 度量方面都优于所有基线模型。结果表明，随着交叉项系数维数的增加，由于基础特征较少，FM 系列模型的表示能力略微上升并逐渐趋于稳定。此外，可以观测到使用了迁移算法的模型比只在目标域学习的模型效果更好。例如，当 K = 3 和测试集实体比例为 10%时，回归情形下 Trans-GLM 模型的平均 RMSE 为 0.9114，比 GLM 模型低 3.3%，而 Trans-FM 模型的平均 RMSE 为 0.8870，比 FM 模型低 3.5%，这说明引入迁移学习算法能有效提升模型的预估能力。整体实验结果说明推荐系统中的冷启动问题在迁移框架的引入下得到了缓解。

在源域选择算法中，我们每一次从 Netflix 数据集中随机采样得到 10 倍于 MovieLens 数据集大小的子集，然后按照评论时间等频切割成 10 份，并从中选择合适的源域。为了直观的评估迁移的效果，我们在下图给出了 K=1 的 Trans-FM 的实验中受选样本的时间分布。

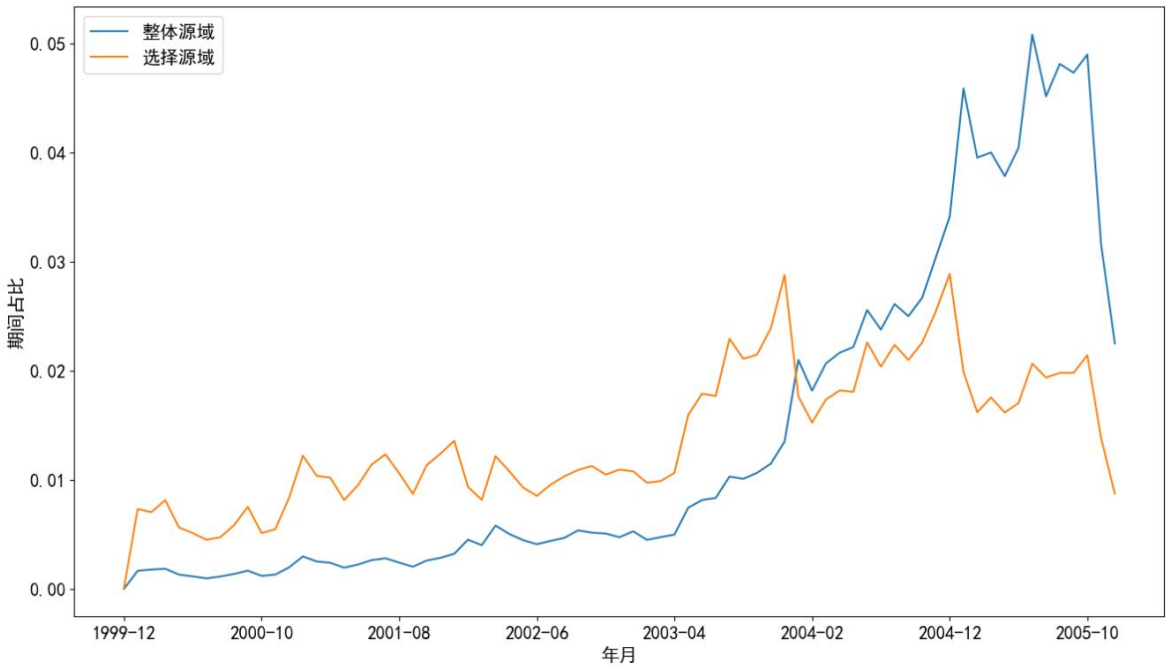

图 4.6 选择前后源域时间分布对比

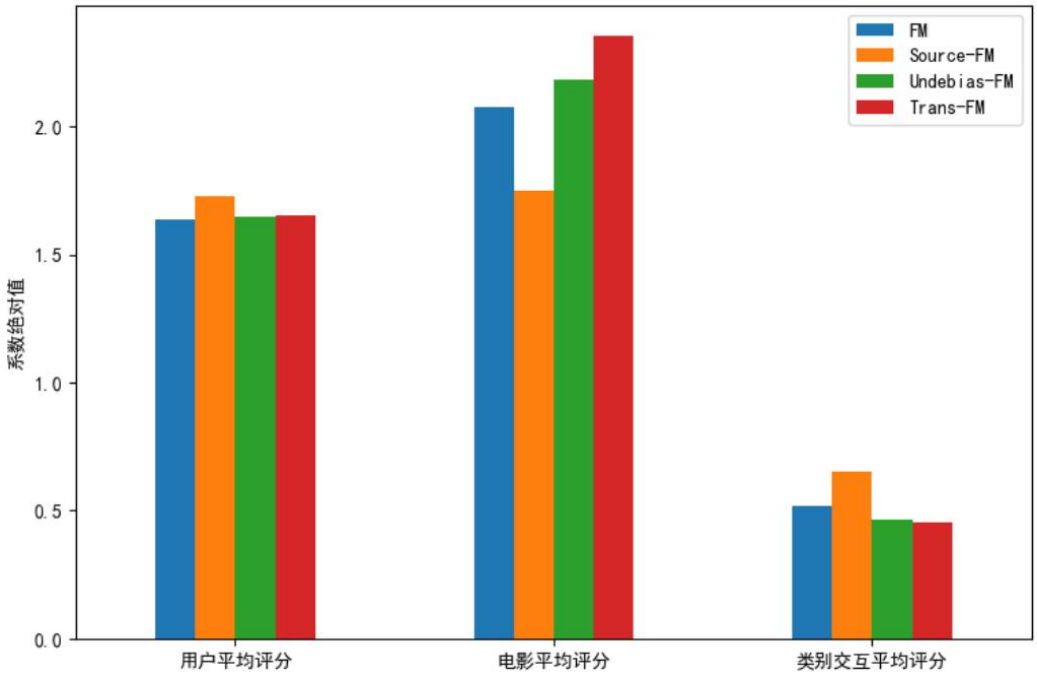

图 4.7 迁移算法各阶段系数对比

从图 4.6 中可以看到，选择算法会偏向于选择更久远的评论，这是由于我们使

用的目标域 MovieLens 数据集是在 2003 年发布的，时间接近的数据迁移效果更好。

除了能在特征稀少的情形下提供较好的预测，线性模型的参数也能提供解释性。在图 4.7 中，我们报告了 K=1 和测试集实体比例为 10%时的一次实验中，Trans-FM 算法在各个阶段的绝对值前三大的参数。图中 FM 代表 FM 只在目标域学习时估计得到的参数，Source-FM 代表 FM 在所有源域学习时估计得到的参数，Undebias-FM 代表选择性迁移但是未纠偏的 FM，Trans-FM 代表所提出模型的全流程得到的估计。从图中可以看到，虽然源域和目标域在一些参数上有较大差异，但是在经历选择性迁移后得到的参数与目标域的参数十分接近，说明选择算法十分有效。测试集上指标的提升也说明了整体迁移算法的有效性，纠偏得到的预估参数更加接近实际情况。

4.3 跨用户冷启动实验

4.3.1 数据集介绍

为了验证模型的跨用户的冷启动效果,我们使用 Douban[53]数据集。在豆瓣平台上，同一个用户能给书籍和电影评分，与形成一个共享用户的跨域推荐场景。可以从 https://github.com/FengZhu-Joey/GA-DTCDR/tree/main/Data 中获取。

Douban 由两部分组成：书籍评分数据、电影评分数据、音乐评分数据。其具体介绍如下：

1.Douban 书籍评分数据：包括用户 ID、书籍 ID、评分。

表 4.11 Douban 书籍数据描述表

	用户 ID	书籍 ID	评分
描述	用户编码	书籍编码	评分标准为仅整数星级的五星级打分
计数	96041	96041	96041
最小值	1	1	5
最大值	2718	6777	5
举例 1	426	3690	4
举例 2	1901	5357	5

2.Douban 电影评分数据：包括用户 ID、电影 ID、评分。

表 4.12 Douban 电影数据描述表

	用户 ID	电影 ID	评分
描述	用户编码	书籍编码	评分标准为仅整数星级的五星级打分
计数	1133420	1133420	1133420
最小值	1	1	1
最大值	2718	9565	5
举例 1	1241	5046	4
举例 2	1421	167	4

3.Douban 音乐评分数据：包括用户 ID、音乐 ID、评分。

表 4.13 Douban 音乐数据描述表

	用户 ID	音乐 ID	评分
描述	用户编码	音乐编码	评分标准为仅整数星级的五星级打分
计数	69709	69709	69709
最小值	2	1	1
最大值	2718	5567	5
举例 1	1507	5196	5
举例 2	2317	4041	4

在用户共享的场景，以音乐评分以及庞大的电影评分为源域，以书籍评分为共享用户的目标域。不同于 MovieLens-Netflix 数据集，Douban 数据集信息更加稀少，只有每个用户在各个数据域的评分。每个用户的兴趣和苛刻程度说明了迁移的可行性，而被评价实体的差异也为迁移算法提出了挑战。

4.3.2 数据预处理

使用的 Douban 数据集统计如下：

表 4.14 Douban 电影-书籍对比表

统计	电影评分	书籍评分
用户数量	2,718	2,718
实体数量	9,555	6,777
评分数量	1,133,420	96,041

不同于 MovieLens-Netflix 数据，Douban 数据集没有提供除了评分外的关于电

影、书籍的任何额外信息，故生成的特征更少。分别在电影评分数据和书籍评分数据上将原始数据加工，加工特征的统计描述如下：

表 4.15 Douban 书籍评分特征表

特征域	用户域			物品域		
特征	评分均值	评分标准差	评分计数	评分均值	评分标准差	评分计数
最小值	1.00	0.00	1.00	1.00	0.00	6.00
中位数	3.96	0.83	111.00	4.00	0.76	17.00
最大值	5.00	2.83	656.00	5.00	1.94	359.00
均值	3.96	0.85	149.32	3.96	0.78	33.12
标准差	0.33	0.20	134.06	0.50	0.21	47.45
举例 1	3.55	1.44	49	3.54	0.82	730
举例 2	3.87	0.83	612	3.86	0.75	397
举例 3	3.68	0.84	1291	4.56	0.72	519

表 4.16 Douban 电影评分特征表

特征域	用户域			物品域		
特征	评分均值	评分标准差	评分计数	评分均值	评分标准差	评分计数
最小值	1.00	0.00	1.00	1.06	0.24	16.00
中位数	3.70	0.91	670.00	3.77	0.80	221.00
最大值	5.00	2.31	3643.00	4.90	1.75	1357.00
均值	3.70	0.92	835.61	3.70	0.81	324.41
标准差	0.31	0.15	624.21	0.55	0.11	294.74
举例 1	3.55	1.44	49	3.54	0.82	730
举例 2	3.87	0.83	612	3.86	0.75	397
举例 3	3.68	0.84	1291	4.56	0.72	519

虽然源域及目标域的数据量差距极大，且评分对象差异很大，但是用户评分分布相近，故有迁移的可行性。

不同于 MovieLens-Netflix 数据集，Douban 数据集没有用户、物品的细节描述，使得子源域的选取较为复杂，而我们提出的 Trans-FM 算法需要依次遍历多个子源域并从中选出合适的迁移数据。由于没有额外信息来划分源域，我们使用 KMeans

聚类算法[54]，用标准化后的特征作为输入，将源域区分为差异较大的子源域，为实现 Trans-FM 算法铺垫。KMeans 聚类是一种迭代的聚类算法，它将数据集划分为 K 个类别。算法首先随机选择 K 个初始聚类中心，然后通过迭代的方式将样本分配到最近的聚类中心，并更新聚类中心位置。该过程在聚类中心不再改变或达到预定的迭代次数后停止。根据肘部法，选择 10 为最终的聚类簇数。对聚类后的数据采样 10000 条并按照区分度较高的特征进行可视化：

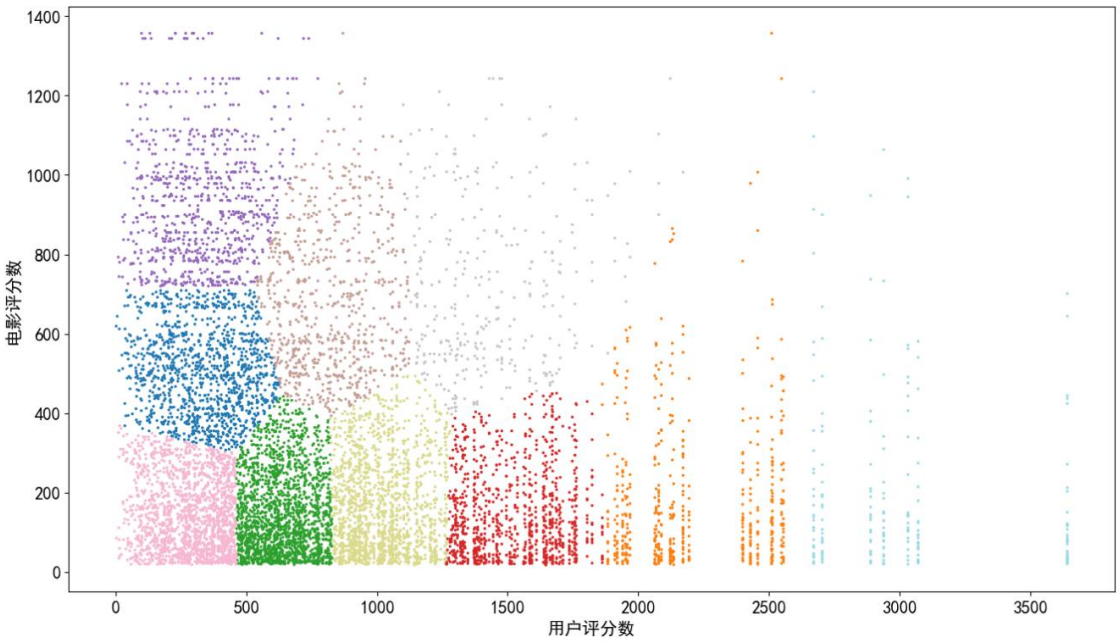

图 4.8 聚类结果特征分布

从图中可以看到部分簇是由评分过大量电影的用户产生的评分，可以认为是专家评分；部分簇是由有大量评分的电影聚成的，可以认为是热品评分。每一簇有一定的聚集性，特征分布也有较大区别，聚类后的每一簇可以当作来自不同的源域。更重要的，由于源域和目标域数据量差距较大，导致在评分数特征上也有较大的差异。需要按照较大差异的特征区分源域，剔除差异较大的数据。

4.3.3 实验结果

在 Douban 数据集上，和 MovieLens-Netflix 数据的设置一样，按一定比例随机删除了目标域中一小部分用户的打分，并将它们作为测试集。并和上一节相同，对测试集设置了不同的占比，即 10%、20%、30%、40%和 50%。各种测试集用户占比设置下的 10 次 5 折交叉验证实验平均结果如下：

表 4.17 跨用户冷启动回归 RMSE 表

		10%	20%	30%	40%	50%
	AVE	0.9187	0.9216	0.9200	0.9158	0.9180
	GLM	0.8073	0.8096	0.8099	0.8101	0.8098
	Trans-GLM	0.7612	0.7620	0.7616	0.7622	0.7625
K=1	FM	0.7395	0.7400	0.7407	0.7410	0.7414
K=1	Trans-FM	0.7090	0.7099	0.7102	0.7105	0.7108
K=2	FM	0.7399	0.7395	0.7400	0.7412	0.7413
K=2	Trans-FM	0.7089	0.7095	0.7105	0.7101	0.7107
K=3	FM	0.7402	0.7397	0.7405	0.7413	0.7410
K=3	Trans-FM	0.7095	0.7103	0.7109	0.7110	0.7111

表 4.18 跨用户冷启动多分类 macro-AUC 表

		10%	20%	30%	40%	50%
	GLM	0.6224	0.6221	0.6219	0.6213	0.6210
	Trans-GLM	0.7233	0.7204	0.7194	0.7189	0.7199
K=1	FM	0.6331	0.6333	0.6320	0.6329	0.6312
K=1	Trans-FM	0.7320	0.7315	0.7317	0.7304	0.7305
K=2	FM	0.6321	0.6343	0.6344	0.6339	0.6333
K=2	Trans-FM	0.7330	0.7323	0.7323	0.7319	0.7311
K=3	FM	0.6324	0.6346	0.6340	0.6343	0.6340
K=3	Trans-FM	0.7328	0.7322	0.7320	0.7315	0.7310

其中所使用的模型和 MovieLens-Netflix 完全相同，区别在于本节是跨用户的冷启动推荐。AVE 推荐方法作为一种简单的方法依旧是性能最差的。所提出的 Trans-FM 模型在 RMSE 和 AUC 度量方面都优于所有基线模型。和 MovieLens-Netflix 中的实验不同，随着交叉项系数维数的增加，FM 系列模型的表示能力没有明显变化，原因是在 Douban 数据集中只有基础评分数据，没有额外的信息，更多的交叉没有使得性能提升。此外，和 MovieLens-Netflix 相同，迁移算法的模型比朴素模型效果更好。例如，当 K = 1 时，Trans-FM 模型的平均 AUC 为 73.122，比 FM 模型的 63.250 高 15.6%。整体实验结果说明 Trans-FM 能有效缓解推荐系统中的冷启动问题。

4.4 信息茧房实验

4.4.1 数据集介绍

用于评估信息茧房的数据集是从电商平台中获取的，叫做 Amazon[55]。Amazon

Review Data（2018）是一个广泛使用的数据集，包含了来自 Amazon 在线商店的大量带时间的产品评价，旨在支持商品评价的情感分析、文本挖掘和推荐系统的研究。Amazon 可以从 https://cseweb.ucsd.edu/~jmcauley/datasets/amazon_v2/中获取。

不同于用于研究冷启动场景的数据，Amazon 数据集提供了商品的原始信息如商品价格、商品类别、商品图像、类似商品集。以及准确的用户评价，如文本评价、打分、评价日期。

在庞大的 Amazon 数据集中，我们只选取了类别为游戏外设和电子游戏的评论数据及商品数据。在信息茧房场景，以电子游戏评论信息为源域，以游戏外设评论信息为共享用户的目标域。从召回率和召回样本多样性两个角度评估模型缓解信息茧房的效果。不同于前两个数据集，在 Amazon 数据集中，我们更关心模型的实践效果：迁移算法能否提供更多样的结果。

4.4.2 数据预处理

在 Amazon 数据集上，我们只使用了游戏外设和电子游戏的评论数据及商品数据，并只保留评论超过 5 次的商品，只保留在源域及目标域评论皆超过 5 次的用户，最终的数据量如下：

表 4.19 Amazon 数据集游戏外设和电子游戏评论数据对比表

统计	游戏外设	电子游戏
用户数量	8,169	8,169
实体数量	14,606	41,565
评分数量	92,611	125,884

在 Amazon 数据集提供了更全面的数据，我们在前 50%时间的训练集上构造特征，并在后 50%测试集上评估召回精度及多样性。不同于前面两个实验，在 Amazon 数据集上我们构造了更加丰富的历史行为特征，源域和目标域的信息将会被同时用于构造特征：为每个用户统计其在源域以及目标域上的整体评论次数。

4.4.3 实验结果

不同于上两个任务，我们按照评论的时间顺序将数据分为前 50%的训练集和后 50%的测试集。在训练集上使用时间序列交叉验证法进行参数选择，利用训练阶段的评论信息，预测用户为物品打分的概率。在训练阶段，将用户评论过的数

据皆作为正样本。没有发生评论的用户-物品对是非常庞大的，所以需要随机采样得到负样本。随机打乱正样本的物品，将没有交互过的物品作为负样本，重复 10 次。最终得到正负样本比例为 1：10。

在 Amazon 数据集上，10 次不同采样种子的实验平均结果如下：

表 4.20 信息茧房召回率表

	Recall@10	Recall@20	Recall
Rule	0.0000	0.0000	0.0008
CF	0.0012	0.0025	0.0301
GLM	0.0039	0.0074	0.0301
Trans-GLM	0.0045	0.0087	0.0301
FM	0.0045	0.0079	0.0301
Trans-FM	0.0051	0.0091	0.0301

表 4.21 信息茧房多样性汉明距离结果表

	Hamming@10	Hamming@20
CF	0.9999	0.9997
GLM	0.9910	0.9934
Trans-GLM	0.9914	0.9940
FM	0.9900	0.9931
Trans-FM	0.9916	0.9950

其中 Rule 是规则召回，代表将用户在训练集中评论过的个物品作为召回结果，CF 是协同过滤，代表将训练集中的用户评论过物品的相似物品作为全部召回结果。其他所有算法的召回结果都是将 CF 的结果重新排序，使得前排的推荐结果更加精准，并选取前 K 个物品计算召回率 Recall@K 及汉明距离 Hamming@K。GLM 为经典线性模型。FM 拓展了交叉项，其交叉表征维数 K 设置为 3。Trans-X 代表迁移算法加持，且只在目标域评估。

容易注意到，所有的召回率皆低于 CF 的整体召回率，这是由于所有的召回结果都是在 CF 之后继续排序选出的子集。Rule 极低的召回率也说明了重复评论发生概率之低，运用算法进行个性化召回十分必要。实验结果显示，所提出的 Trans-FM 模型在召回精度及召回结果的多样性方面都优于基线模型，且迁移学习算法能够提高推荐结果的多样性。

第五章 结论

5.1 工作总结

近年来，推荐系统在互联网行业得到了广泛的应用和发展。随着人工智能和机器学习技术的不断发展，推荐系统的算法越来越智能化和个性化，能够更准确地预测用户的喜好。随着技术的不断发展，跨域推荐成为当前推荐系统研究中一个新的研究热点。面对多领域数据，如何有效的利用信息并针对主要目标建模被学界业界广泛研究并应用。但冷启动与信息茧房等问题依然存在，仍然是相关领域研究的重点。

本文基于上述研究背景对跨域推荐的统计推荐模型展开研究，总结了目前已有研究所取得的进展，高度概括了统计模型在推荐领域的历史发展以及跨域推荐的近期研究进展。具体成果与结论如下：

1.本文提出了一个迁移学习框架下的因子分解机模型，旨在解决冷启动与信息茧房问题。基模型选取经典的因子分解机模型，借用模型的可解释性以及稳定的优点来缓解冷启动与信息茧房。再将模型引入跨域推荐框架下，先在源域选取合适的样本，再混合目标域样本共同训练得到模型参数的初步估计，最终使用纠偏技巧缓解模型的负迁移现象。使得模型能够在多域场景下得到更有效的结果。

2.本文选取三个跨域推荐领域常见的数据集对模型的推断效果进行了评估。选取RMSE及AUC指标评价回归及分类情形下模型推荐效果以及模型对冷启动的缓解程度，选取召回率及汉明距离评价模型对信息茧房的缓解程度。通过三个不同的实验证明了，所提出的模型在跨物品冷启动场景、跨用户冷启动场景、信息茧房场景下对传统模型有显著提升。

5.2 工作展望

本文提出了一种利用跨域推荐提升统计推荐模型的可行方案，但在诸多方面仍存在不足，后续研究可在本文的基础之上加以改进。

1.在跨领域信息的利用上，源域信息的特征需要和目标域信息的特征完全重合。但是在实际的工业推荐场景中，会出现两者信息不重叠的部分，利用非重叠部分

进行信息的增强能够更进一步提高推荐的效果。

2.在基模型的选取上，本文在推荐领域常见模型中选取了可解释很强的因子分解机模型，以便于对跨域推荐进行可解释性分析。但是，本文所使用的迁移学习算法可以推向更宽泛的模型，如 FFM 统计推荐模型、Wide&Deep 深度学习模型。该框架能够帮助大部分单域模型利用多领域的信息。

参考文献

[1] Olivas E S, Guerrero J D M, Martinez-Sober M, et al. Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods, and Techniques: Algorithms, Methods, and Techniques [M]. IGI Global, 2009.

[2] Zhu F, Chen C, Wang Y, et al. Dtcdr: A Framework for Dual-Target Cross-Domain Recommendation. Proceedings of the 28th ACM International Conference on Information and Knowledge Management, F, 2019 [C].

[3] Kumar B, Sharma N. Approaches, Issues and Challenges in Recommender Systems: A Systematic Review [J]. Indian Journal of Science & Technology, 2016, 9(47): 1-12.

[4] Nguyen T T, Hui P-M, Harper F M, et al. Exploring the Filter Bubble: The Effect of Using Recommender Systems on Content Diversity. Proceedings of the 23rd International Conference on World Wide Web, F, 2014 [C].

[5] Schein A I, Popescul A, Ungar L H, et al. Methods and Metrics for Cold-Start Recommendations. Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, F, 2002 [C].

[6] Li S, Cai T T, Li H. Transfer Learning for High-Dimensional Linear Regression: Prediction, Estimation and Minimax Optimality [J]. Journal of the Royal Statistical Society Series B, 2022, 84(1): 149-173.

[7] Resnick P, Varian H R. Recommender Systems [J]. Communications of the ACM, 1997, 40(3): 56–58.

[8] Linden G, Smith B, J Y. Amazon.Com Recommendations: Item-to-Item Collaborative Filtering [J]. Internet Computing, IEEE, 2003, 7(1): 76-80.

[9] Hosmer Jr D W, Lemeshow S, Sturdivant R X. Applied Logistic Regression [M]. John Wiley & Sons, 2013.

[10] Baraldi A, Panniggiani F. An Investigation of the Textural Characteristics Associated with Gray Level Cooccurrence Matrix Statistical Parameters [J]. IEEE Transactions on Geoscience and Remote Sensing, 1995, 33(2): 293-304.

[11] Breese J S. Empirical Analysis of Predictive Algorithms for Collaborative Filtering. Proceedings of the 14th Conference on Uncertainty in Artificial Intelligence, Madison, WI, F, 1998 [C].

[12] Koren Y, Bell R, Volinsky C. Matrix Factorization Techniques for Recommender Systems [J]. Computer, 2009, 42(8): 30-37.

[13] Agarwal D, Chen B-C. Regression-Based Latent Factor Models. Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, F, 2009 [C].

[14] Rendle S. Factorization Machines. 2010 IEEE International Conference on Data Mining, F, 2010 [C]. IEEE.

[15] Juan Y, Zhuang Y, Chin W-S, et al. Field-Aware Factorization Machines for Ctr Prediction. Proceedings of the 10th ACM Conference on Recommender Systems, F, 2016 [C].

[16] Pan J, Xu J, Ruiz A L, et al. Field-Weighted Factorization Machines for Click-through Rate Prediction in Display Advertising. Proceedings of the 2018 World Wide Web Conference, F, 2018 [C].

[17] Pande H. Field-Embedded Factorization Machines for Click-through Rate Prediction [J]. arXiv preprint arXiv:200909931, 2020.

[18] Cheng H-T, Koc L, Harmsen J, et al. Wide & Deep Learning for Recommender Systems. Proceedings of the 1st Workshop on Deep Learning for Recommender Systems, F, 2016 [C].

[19] Guo H, Tang R, Ye Y, et al. Deepfm: A Factorization-Machine Based Neural Network for Ctr Prediction [J]. arXiv preprint arXiv:170304247, 2017.

[20] Wang R, Fu B, Fu G, et al. Deep & Cross Network for Ad Click Predictions [M]. Proceedings of the Adkdd'17. 2017: 1-7.

[21] Wang R, Shivanna R, Cheng D, et al. Dcn V2: Improved Deep & Cross Network and Practical Lessons for Web-Scale Learning to Rank Systems. Proceedings of the Web Conference 2021, F, 2021 [C].

[22] Lian J, Zhou X, Zhang F, et al. Xdeepfm: Combining Explicit and Implicit Feature Interactions for Recommender Systems. Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, F, 2018 [C].

[23] Gao Y, Yang Y. Transfer Learning on Stratified Data: Joint Estimation Transferred from Strata [J]. Pattern Recognition, 2023, 140: 109535.

[24] Bastani H. Predicting with Proxies: Transfer Learning in High Dimension [J]. Management Science, 2021, 67(5): 2964-2984.

[25] Tian Y, Feng Y. Transfer Learning under High-Dimensional Generalized Linear Models [J]. Journal of the American Statistical Association, 2023, 118(544): 2684-2697.

[26] Li S, Zhang L, Cai T T, et al. Estimation and Inference for High-Dimensional Generalized Linear Models with Knowledge Transfer [J]. Journal of the American Statistical Association, 2023: 1-12.

[27] Candes E, Tao T. The Dantzig Selector: Statistical Estimation When P Is Much Larger Than N [J]. The Annals of Statistics, 2007, 35(6): 2313-2351.

[28] Pan W, Xiang E, Liu N, et al. Transfer Learning in Collaborative Filtering for Sparsity Reduction. Proceedings of the AAAI Conference on Artificial Intelligence, F, 2010 [C].

[29] Man T, Shen H, Jin X, et al. Cross-Domain Recommendation: An Embedding and Mapping Approach. IJCAI, F, 2017 [C].

[30] Zhu F, Wang Y, Chen C, et al. A Deep Framework for Cross-Domain and Cross-System Recommendations [J]. arXiv preprint arXiv:200906215, 2020.

[31] Kang S, Hwang J, Lee D, et al. Semi-Supervised Learning for Cross-Domain Recommendation to Cold-Start Users. Proceedings of the 28th ACM International Conference on Information and Knowledge Management, F, 2019 [C].

[32] Zhu Y, Ge K, Zhuang F, et al. Transfer-Meta Framework for Cross-Domain Recommendation to Cold-Start Users. Proceedings of the 44th International

ACM SIGIR Conference on Research and Development in Information Retrieval, F, 2021 [C].

[33] Singh A P, Gordon G J. Relational Learning Via Collective Matrix Factorization. Acm Sigkdd International Conference on Knowledge Discovery & Data Mining, F, 2008 [C].

[34] Jiang M, Cui P, Yuan N J, et al. Little Is Much: Bridging Cross-Platform Behaviors through Overlapped Crowds. Proceedings of the AAAI Conference on Artificial Intelligence, F, 2016 [C].

[35] Hu G, Zhang Y, Yang Q. Conet: Collaborative Cross Networks for Cross-Domain Recommendation. Proceedings of the 27th ACM International Conference on Information and Knowledge Management, F, 2018 [C].

[36] Cui Q, Wei T, Zhang Y, et al. Herograph: A Heterogeneous Graph Framework for Multi-Target Cross-Domain Recommendation. ORSUM@ RecSys, F, 2020 [C].

[37] Zhang W, Zhang P, Zhang B, et al. A Collaborative Transfer Learning Framework for Cross-Domain Recommendation. Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, F, 2023 [C].

[38] Ostertagova E. Modelling Using Polynomial Regression [J]. Procedia Engineering, 2012, 48: 500-506.

[39] Pan S J, Tsang I W, Kwok J T, et al. Domain Adaptation Via Transfer Component Analysis [J]. IEEE Transactions on Neural Networks, 2010, 22(2): 199-210.

[40] Zhang Y, Yang Q. An Overview of Multi-Task Learning [J]. National Science Review, 2018, 5(1): 30-43.

[41] Mikolov T, Sutskever I, Chen K, et al. Distributed Representations of Words and Phrases and Their Compositionality [J]. Advances in Neural Information Processing Systems, 2013, 26.

[42] Ben-David S, Blitzer J, Crammer K, et al. A Theory of Learning from Different Domains [J]. Machine Learning, 2010, 79: 151-175.

[43] Bobadilla J, Ortega F, Hernando A, et al. A Collaborative Filtering Approach to Mitigate the New User Cold Start Problem [J]. Knowledge-Based Systems, 2012, 26: 225-238.

[44] Zhu Y, Lin J, He S, et al. Addressing the Item Cold-Start Problem by Attribute-Driven Active Learning [J]. IEEE Transactions on Knowledge and Data Engineering, 2019, 32(4): 631-644.

[45] Cohn D A, Ghahramani Z, Jordan M I. Active Learning with Statistical Models [J]. Journal of Artificial Intelligence Research, 1996, 4: 129-145.

[46] Pariser E. The Filter Bubble: What the Internet Is Hiding from You [M]. Penguin UK, 2011.

[47] Konstan J A, Miller B N, Maltz D, et al. Grouplens: Applying Collaborative Filtering to Usenet News [J]. Communications of the ACM, 1997, 40(3): 77-87.

[48] McCullagh P. Generalized Linear Models [M]. Routledge, 2019.

[49] Kingma D P, Ba J. Adam: A Method for Stochastic Optimization [J]. arXiv preprint arXiv:14126980, 2014.

[50] Pan S J, Yang Q. A Survey on Transfer Learning [J]. IEEE Transactions on Knowledge and Data Engineering, 2009, 22(10): 1345-1359.

[51] Prokhorenkova L, Gusev G, Vorobev A, et al. Catboost: Unbiased Boosting with Categorical Features [J]. Advances in Neural Information Processing Systems, 2018, 31: 6638-6648.

[52] Mangalathu S, Hwang S-H, Jeon J-S. Failure Mode and Effects Analysis of Rc Members Based on Machine-Learning-Based Shapley Additive Explanations (Shap) Approach [J]. Engineering Structures, 2020, 219: 110927.

[53] Huang J, Cheng X, Shen H, et al. Exploring Social Influence Via Posterior Effect of Word-of-Mouth Recommendations. Proceedings of the 5th International Conference on Web Search and Web Data Mining, WSDM 2012, Seattle, WA, USA, February 8-12, 2012, F, 2012 [C].

[54] Friedman H P, Rubin J. On Some Invariant Criteria for Grouping Data [J]. Journal of the American Statistical Association, 1967, 62(320): 1159-1178.

[55] Fu W, Peng Z, Wang S, et al. Deeply Fusing Reviews and Contents for Cold Start Users in Cross-Domain Recommendation Systems. Proceedings of the AAAI Conference on Artificial Intelligence, F, 2019 [C].

[56] 许海玲, 吴潇, 李晓东等. 互联网推荐系统比较研究 [J]. 软件学报, 2009, 20(02): 350-362.

[57] 吴正洋, 汤庸, 刘海. 个性化学习推荐研究综述 [J]. 计算机科学与探索, 2022, 16(01): 21-40.

[58] 罗浩, 高升, 徐蔚然. 基于个性信息的跨域推荐算法 [J]. 软件, 2013, 34(12): 142-147.

[59] 陈雷慧, 匡俊, 陈辉等. 跨领域推荐技术综述 [J]. 华东师范大学学报(自然科学版), 2017, (05): 101-116+137.