***MPMA*: Mixture Probabilistic Matrix Approximation for Collaborative Filtering**

# Abstract

矩阵近似（MA）是协作过滤（CF）最流行的技术之一。大多数现有MA方法基于用户项目评级矩阵训练用户/项目延迟因子，然后使用全局潜在因子来模拟所有用户/项目。然而，全局优化的潜在因素可能不反映仅在用户/项目的子集之间共享的独特兴趣，否则可能无法准确地建模用户的独特兴趣。因此，现有的MA方法无法捕获不同用户/项目的唯一性，无法提供最佳推荐。

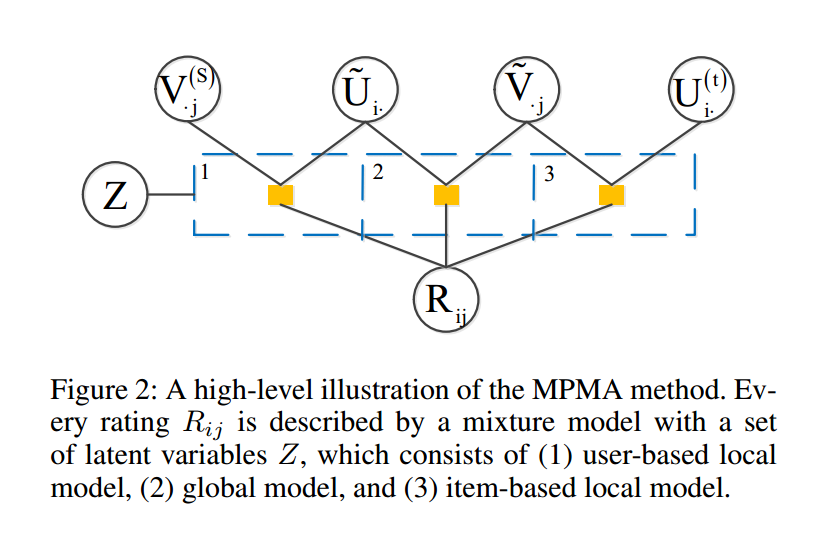
本文提出了一种混合概率矩阵近似（MPMA）方法，该方法统一了全局优化用户/项目特征向量（在整个评级矩阵上）和局部优化的用户/项目特征向量（在用户/项目评级的子集上），以提高建议的准确性。更具体地，在MPMA中，开发了一种方法来找到全局和局部优化的用户/项目特征向量。然后，采用高斯混合模型将全局预测和局部预测结合起来，以产生准确的评级预测。使用MovieLens和Netflix数据集的实验研究表明，MPMA在推荐精度和优秀的可扩展性方面优于五种最先进的基于MA的CF方法。

# INTRODUCTION

协同过滤（CF）方法在当今的推荐系统中取得了巨大的成功，其中矩阵近似（MA）是最流行的技术之一。在基于MA的CF方法中，用户和项目的特征是**从用户项目评级矩阵推断出的潜在因素向量，并且这些潜在因素用于进行评级预测**[Adomavicius和Tuzhilin，2005; Su和Khoshgoftaar，2009] 。大多数现有的基于MA的方法[Srebro et al，2004; Salakhutdinov和Mnih，2007年; Koren，2008]依靠全局优化的用户/项目潜在因素来提出建议。然而，在许多现实世界的应用中，**如果我们将全局优化的潜在因素视为“共同兴趣”，那么用户子集可能会分享“共同兴趣”未涵盖的“独特兴趣”**[Xu et al，2012]。因此，如果MA方法仅考虑“共同兴趣”而不考虑“独特兴趣”，则推荐对于每个用户的“唯一”子集可能不是最佳的。正如Koren等人[2008]所指出的，MA模型可以有效地估计与大多数或所有项目同时相关的整体结构，但是它们在检测一小组项目之间的强关联方面表现不佳。

最近，**矩阵聚类[Xu et al，2012; Lee等人，2013]和社区检测[Zhang et al，2013]方法已被提出用于发现用户/项目的子集之间的局部化关系。在这些方法中，在群集/社区内训练局部用户/项目潜在因素，然后由局部模型产生推荐。然而，这些方法通常仅依赖于局部模型，并且未能将全局潜在因素纳入推荐中，由于局部集群或社区中的数据不足，这可能会影响推荐质量[Chen et al，2015]**。总之，基于MA的CF方法仅依赖于全局潜在因素或局部潜在因素而损害推荐准确性。因此，需要设计新的基于MA的CF方法，其可以结合全局潜在因子和局部潜在因子。

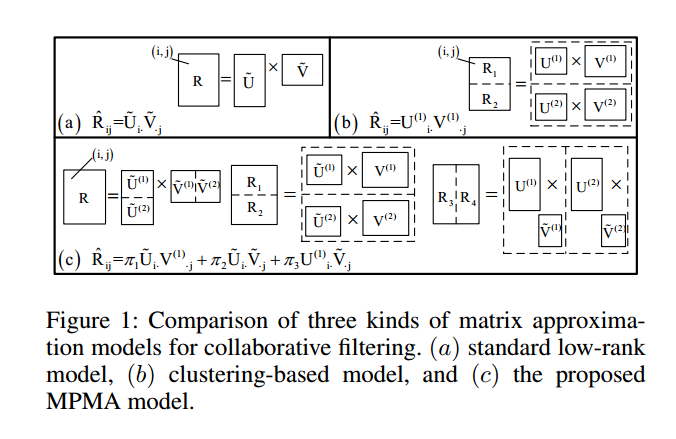
在本文中，我们开发了一种新方法，使用局部和全局潜在因素对每个用户/项目进行建模，以捕获用户项子组中的局部化关系以及所有用户和项目之间的公共关联。遵循多任务特征学习技术[Evgeniou和Pontil，2007; Ando和Zhang，2005]，我们在用户项子组中共享全局用户/项目潜在因素，以便可以训练局部潜在因素而不会出现数据不足的问题。在贝叶斯视角中，所提出的混合概率矩阵近似（MPMA）方法假设每个用户项目评级可以由**包含三个组件的高斯混合模型描绘**：（1）捕获所有用户和项目的全局潜在因素的全局模型（2）基于用户的局部模型，其捕获相应用户子集的局部潜在因素，以及（3）基于项目的局部模型，其捕获相应项目子集的局部潜在因素。在MPMA中，首先定义并解决新的优化问题以获得每个用户/项目的全局/局部潜在因子。提出了一种基于流水线的学习方法，用于并行训练用户/项目潜在因素。最后，通过组合来自混合模型的不同组分的分数来生成推荐分数。使用两个真实数据集（MovieLens和Netflix）评估所提出的MPMA方法，并且实验结果表明MPMA比五种最先进的基于MA的CF方法实现更好的推荐精度，同时实现了良好的效率。



# 2 Related Work

推荐系统近年来越来越受欢迎，旨在为符合用户口味的产品提供个性化推荐[Adomavicius和Tuzhilin，2005]。在现有的推荐器解决方案中，协同过滤（CF）被广泛用于实现的简单性和高质量的推荐。早期的CF算法专注于基于记忆的方法，例如基于用户的[Herlocker等人，1999]和基于项目的[Sarwar等人，2001]方法，其基于用户/项目之间的相似性进行评级预测。然而，这些方法遭受数据稀疏性问题，**因为如果没有足够的评级，则无法准确地计算用户/项目的相似性。**

最近，已经提出基于矩阵近似的CF方法来减轻数据稀疏性问题。Billsus等人[1998]首先将SVD引入协同过滤领域。后来，Srebro等人[2004]提出了最大边际矩阵分解（MMMF）方法。Salakhutdinov等人[2007]首先提出了概率矩阵分解（PMF）方法，后来构建了BPMF - PMF方法的贝叶斯扩展[Salakhutdinov和Mnih，2008]。Koren等人[2008]指出这些模型通常有效地估计与大多数或所有项目同时相关的整体结构，但在检测一小组密切相关项目之间的强关联方面表现不佳。为了解决这个问题，最近的工作采用了矩阵聚类技术[Xu et al。，2012]和社区检测方法[Zhang et al。，2013]来找到具有强相关性的用户/项目聚类，以提高推荐准确性。Mackey等人[2011]和lee等人[2013]遵循分而治之的方法，将MA任务划分为较小的子问题，并行解决这些子问题，然后结合子模型的建议以实现更高的准确性。但是，**这些方法主要关注集群内的评级，忽略集群外的大多数用户评级。由于在检测到的群集中训练数据通常不足，因此严重过度拟合会导致局部模型的性能下降[Chen et al。，2015]。**图1分别总结了（a）和（b）中这两种类型的MA模型。



这项工作的目标是统一用户项子组中的局部化关系以及所有用户和项目之间的共同关联，以提高推荐准确性。最相关的现有工作是集体矩阵分解（CMF）[Singh和Gordon，2008]和Group-Sparse Matrix Factorization（GSMF）[Yuan et al。，2014]。当分解为多个关系表示的多个矩阵时，CMF在因子之间共享参数以学习不同类型的用户行为。GSMF使用组稀疏正则化在多种行为之间自动传递信息。与以前的工作不同，所提出的工作仅利用评级矩阵而不需要额外的信息，并且通过全局（公共）和局部（私有）特征直接对每个用户/项目建模，然后跨多个任务共享全局特征。

我们在图1（c）中说明了这个想法，我们可以看到，R1中的用户项目由全局特征V~和局部特征V（1）建模，全局用户特征U~（1）和U~（2）在多个任务之间共享。值得注意的是，所提出的MPMA方法可以同时学习多个相关任务，因为多任务特征学习已经凭经验和理论证明相对于独立学习每个任务经常显着提高模型性能[Evgeniou和Pontil，2007; Ando和Zhang，2005]。为了更好地理解，我们以贝叶斯的角度介绍了提出的MPMA方法。