原文：Social Recommendation with Missing Not at Random Data

--Jiawei Chen, Can Wang, Martin Ester, Qihao Shi, Yan Feng, and Chun Chen

出自icdm（international conference on data mining）2018

基于缺失非随机假设的社会推荐

**摘要：**

随着在线社交网络使用的爆炸式增长，（学者们）提出了很多社交推荐算法并且证明了社交信息可以被用来提升推荐效果。但是现存的社交推荐方法总是假设数据是随机丢失的（MAR），但是现实生活中很少存在这种情况。我们分析了两个现实的社交推荐数据集，发现了以下两个现象：

（1）用户通常会倾向于选择消费他们喜欢的或者他们朋友购买过的商品

（2）对于一个用户，更多朋友消费该商品时，物品的评分会低于现有模型预测的分数。

为了对这些现象进行建模，我们在社交推荐中引入MNAR假设，并且提出一个新的社交推荐模型SPMF-MNAR，这个对用户评分过程建模。在现实大规模数据集上的实验验证了SPMF-MNAR比现存其它模型获得了更优的效果。

**介绍：**

现今推荐系统已经成为很多在线服务的提供商：比如Amazon和IMBD，的一个核心组成部分。协同过滤是非常流行的推荐算法模型，通过用户的历史反馈预测用户喜好和生成推荐。但是现实生活中，人们只会消费或者评分很少部分的物品，因此传统CF算法面临着稀疏问题，为了解决这个问题，很多算法提出在推荐系统中引入社交信息，这些方法通过对社交关系对用户评分的影响去证明基于社交的协同过滤模型有改善推荐效果的潜力和可能性。

但是，现存社交推荐算法统采用信息随机丢失（MAR）假设，也就是说用户购买物品的过程（或用户评分过程）是与用户评分是无关的。事实上，MAR假设不适用于推荐数据，因为用户通常会购买他们喜欢的东西。相关的，通常高评分通常在评分矩阵中，低评分经常缺失。当MAR假设不成立时，之前的社交推荐算法在预测上就存在一定偏差，并且导致推荐效果受损。

为了更好的理解在社交推荐中评分数据的产生过程，在这篇论文中我们对两个知名的社交推荐数据集做了一个系统的分析：

（1）用户倾向于消费和评分他们喜欢的物品。

（2）用户倾向于购买他们朋友购买过的物品。

（3）社交影响对用户购买情况的影响大于对评分的影响。

事实上，当一个物品被我们更多朋友购买的时候，物品的实际评分比现在的模型预测分数要低。理由如下：在社交影响下，我们有时候会买一些不符合我们口味的东西，因此，这些物品的评分会低于其它消费品。

根据观察结果，在这篇论文中在社交推荐中引入缺失非随机假设（MNAR假设），并且提出了一种新的推荐算法（SPMF-MNAR）。SPMF-MNAR模型对评分观察过程和评分同时建模。如图1所示，一方面SPMF-MNAR根据社交影响和他们个人喜好预测用户的消费情况（也就是评分观察过程），另一方面采用传统协同过滤模型，根据用户喜好去预测评分。特别的，和其他专注于探究社交影响对用户评分影响的预测模型不同，此论文专注于探究社交影响对是否会评分的影响过程。本文使用SVI（即随机变分intergerence ，stochastic variational intergerence）接口去加速我们的模型。我们的实验证明了，我们的模型的有效性。

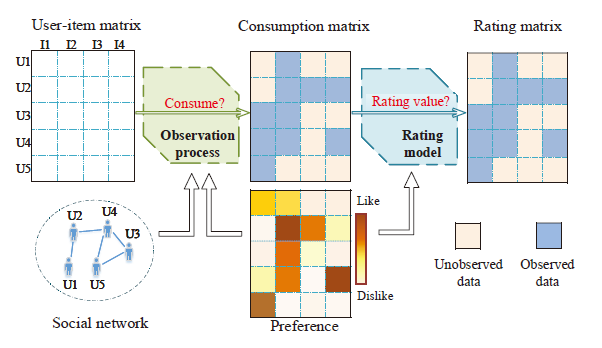


图1

这篇论文的主要贡献如下：

我们提出了在社交推荐中，缺失数据不是随机的思想，并在接下来的实验中在真实数据集上探究了以下几个问题：

（1）在社交推荐中应用MNAR模型是否有好处？

（2）哪些因素会促进用户消费？

（3）提出了一种新的社交推荐的概率模型SPMF-MNAR，这个模型通过把基于社交关系生成的评分过程融入经典的协同过滤评分预测模型中，此外，我们建立以一个高效的随机变量法去推理我们后面的概率模型。

我们通过大型现实数据集评估，证明了我们的方法基于MNAR假设在社交推荐方法和非社交方法上都优于现在的最优推荐算法。

本篇论文的剩下部分将讲述以下内容：在第二部分简要概述了相关工作，第三部分将讲述社交实验分析，第四部分我们展示了SPMF-MNAR模型的细节，第五部分展示了实验结果和分析。论文的最后我们总结了全文，并且提供了未来工作的方向。

**相关工作：**

随着用户评论信息的指数性增长，推荐系统在学术和工业领域越来越受关注。对于有显示反馈的系统（例如评分系统），近二十年不断针对协同过滤准确性和规模性的研究从未中断。因此。我们回顾和我们研究内容最相关的两个领域的内容：一方面是社交推荐相关的，另一方面是基于MNAR假设的推荐。

**社交推荐：** 为了解决传统CF模型冷启动的局限问题，很多通过把社交信息融入推荐系统的模型被提出。这些方法的基本假设是相连接的用户都有类似的喜好，比如，例如一些社交推荐方法包括：SoRec，TrustSVD，PSLF，把评分矩阵和社交（信任）矩阵在共享同一个潜在的用户空间后联合分解。后来将SoRec模型扩展到他们的TrustMF混合模型中。TrustMF混合模型包括了一个信任者模型和一个被信任者模型。他们相信无论活跃用户是信任者还是被信任者都会影响其余用户的评分。此外，一些方法归一化了社交网络中的规则去研究社交关系对评分的影响。特别的，SocialMF介绍了一种新的规则项去限制潜在的喜好靠近他信任的朋友的均值。在其他方法中，用户评分都考虑了用户喜好和社交影响的混合结果。特别的，RSTE模型中，用户评分的数值是通过用户自己的喜好和朋友的喜好共同生成的

但是现有社交推荐算法通常都假设评分数据是随机丢失的（MAR）。事实上，我们在第三部分的分析中，证明了这个基本假设是不对的。隐私，在没有对评分过程观察的情况下，对评分的推断是有偏差的，而且准确率会受损。此外，我们观察了社会关系对用户购买行为的影响比对用户评分的评分的影响更大。所以对于在观察用户评分行为（即用户购买过程）会是有趣并有效的，但是并没有人研究这个工作。

**基于MNAR假设的推荐：** 最近大多数的CF方法都假设CF数据是随机丢失的。也就是说，选择观察数据的过程是和未观察数据值是无关的。但是，就想论文中提到的一样，当MAR假设不成立时，所有的推断都是有偏差的而且预测效果会受损。事实上，有很多证据表明CF数据缺失不是随机的（MNAR）。因此，为了解决这个问题，一些机遇MNAR假设的CF算法在近十年被提出。这些方法生成了了用户的消费过程去对用户是否会购买（评分）以及用户评分进行建模。Marlin 和Zemel通过混合多项式（MM）模型去预测用户评分并且根据评分值推测消费。Hernandez-Lobato 等人观察到混合多项式（MM）模型灵活性太差，并且发现了矩阵分解的好处，因为他们概率矩阵分解模型区处理不是随机丢失的（MNAR）CF数据。最新的，Ohsawa等人在研究了用户消费东西的原因和用户评分之间的依赖性后又把PMF中的模型发展为GPMF模型。因此，用户的消费情况被添加到GPMF模型中去控制不同潜在维度在评分矩阵上进行概率矩阵分解的权重。

但是，所有的方法都忽略社交关系对用户行为的影响。事实上，在社交网络链接的用户通常会有意无意相会推荐商品，所以社交关系毫无疑问会影响他们的消费行为。因此，本文提出SPMF-MNAR模型，对社交影响对用户消费行为建模，从而更好的研究用户对物品的喜好。

**在社交推荐数据集中的用户消费行为：**

在这个部分，我们对两个基于社交推荐的数据集Epinions和Ciao做了一个全面的分析。我们的分析主要是为了解决以下几个问题：

（1）在社交推荐中引入MNAR假设是有利的么？

（2）哪些原因促进了用户的消费？

**数据集介绍：**我们采用了Epinions和Ciao数据集作为我们社交推荐分析和实验的数据。Epinion和Ciao是两个知识分享的网站，用户在网站里对他们消费过的物品进行评分，评分区间在1～5之间。这些数据集报好了以下几种信息：消费信息—一个用户消费了哪些商品，因为用户在评分之前必须消费这个商品，所以这应该是观察评分过程的一部分。还有评分信息，意味着用户给他消费过的物品的评分值。此外，这些数据集还包含了社交信息，每个用户在网站中都有一个朋友列表，这个列表可以直接用来构建一个用户的社交网络。因此，Epinions和Ciao数据集是社交推荐分析的理想数据集，所以被广泛使用，数据集的统计信息被展现在表1中。

|  |  |  |
| --- | --- | --- |
| 特征 | Epinion | Ciao |
| 用户数 | 49,289 | 7,375 |
| 项目数 | 139,738 | 106,797 |
| 评分数 | 664,824 | 280,391 |
| 评分密度 | 0.0097% | 0.0356% |
| 边数 | 487,183 | 111,781 |
| 边密度 | 0.0201% | 0.2055% |

表1

为了解决上述问题，我们做了五个统计分析。（1）我们计算了这些被消费的物品的评分情况的分布，如图2（a）所示（2）我们计算了链接用户的评分相似性和消费相似性，评分相似性系数用pearson系数评估，消费相似性用jaccard系数评估。为了做对比，我们还计算了两个随机用户之间的评分相似性和消费相似性的均值，结果都被展现在图2（d）（e）中。（3）我们我们根据用户的共同好友数，把他们划分为不同的小组。我们计算了每个小组的平均评分相似性和消费相似性，结果被展示在图2（f）（g）中。（4）我们把观察到的评分根据有多少个好友消费过它划分成不同的小组，然后我们计算了平均评分和低分率（低于3分）结果被展示到图2（b）（c）中。

从结果中，我们可以观察出4个重要的结论。

观察结果1:用户倾向于消费（评分）他们喜欢的物品：如图2（a）所示，超过70%的评分大于3，只有少于15%的数据小于3.用户对他们消费过的商品打高分，意味着用户倾向于购买他们认为他们会喜欢的物品，这个结论和论文4中的发现是相同的。他们在线做了一个测试去收集对于随机商品用户的评分，然后发现用户对于随机商品的评分是远低于用户对于他们自己消费过的商品的评分。

观察结果2:用户倾向于消费（评分）他们的朋友消费过的商品

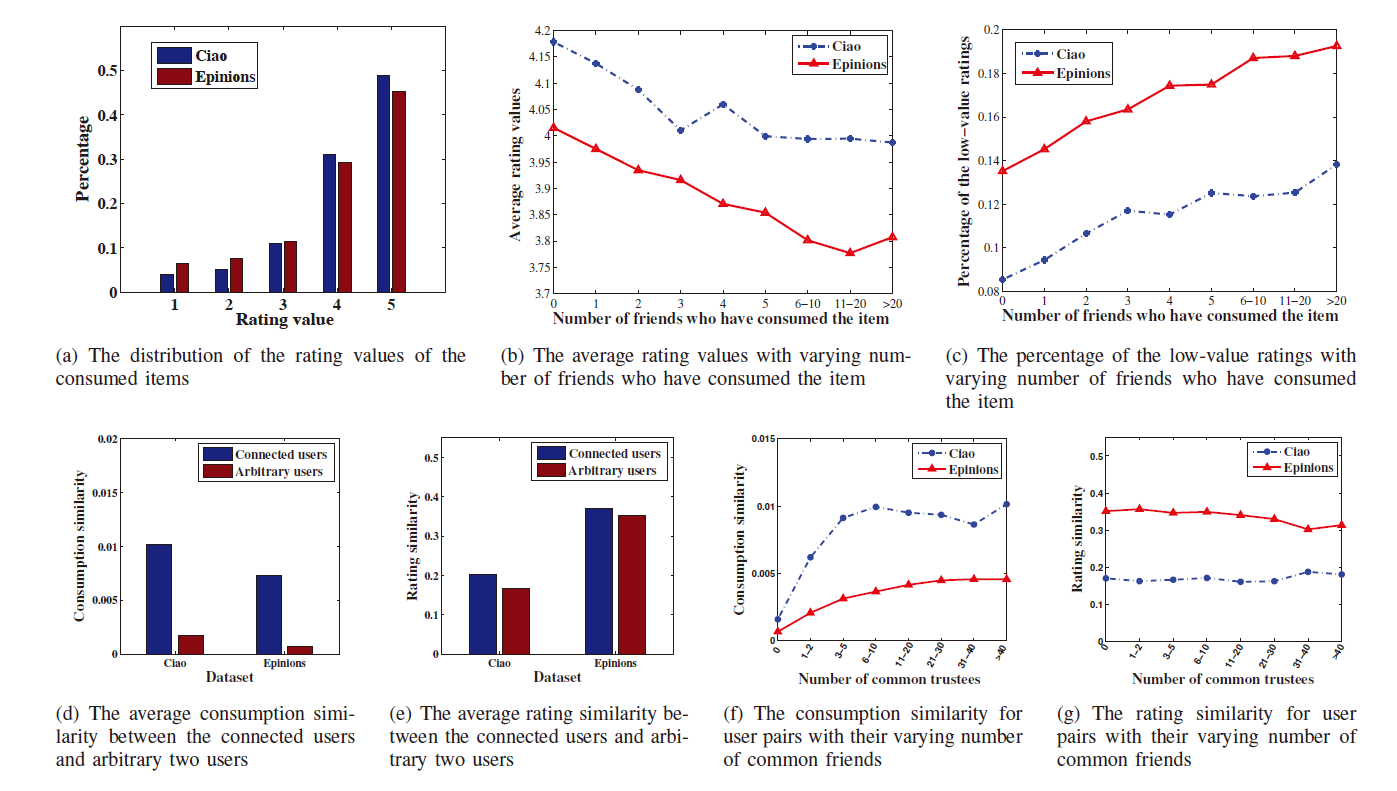
观察结果3:朋友对用户消费的影响大于对评分的影响：如图2（d）（e）所示，链接用户通常会消费更多的相同物品，但是我们的评分并没有比随机用户组合对于他们的消费评分有更多的相似性。与此同时，随着共同好友数量的增加，用户消费的商品会有更高的重合性，但是他们的评分值并没有随着他们共享关系的增加而增加。因此，和用户评分相比，用户消费情况更容易受到社交关系的影响。

观察结果4:消费了这个商品的朋友越多，用户对该商品的评分越低，如图2（b）（c）所示，当物品被更多朋友消费之后，用户的评分越低，并且低分率越高。

基于以上观察结果，我们可以回答上面提到了两个问题。

**结论1:**在社交推荐中应用MNAR模型是有益的。一方面，基于我们的观察，我们可以得出结论评分数据不是随机丢失的。特别的，用户倾向于消费他们喜欢的物品，并且低评分通常是缺失项，基于观察结果1.当我们处理MNAR数据但是并不基于MNAR的假设，推断会变差，预测的准确性也会受损。另一方面，就如同观察结果3中所提到的，用户消费更容易收到社交关系影响。所以说研究社交关系对消费的影响可以让我们更准确的研究用户真实的喜好，

**结论2:**用户的消费情况可以认为收到用户个人喜好和社交影响双方面的影响：一方面，喜好应该发生在选择之前。当用户决定要不要购买某个商品时，他们会衡量他们时候会喜欢这个商品.另一方面，在社交网络中，我们都属于内容共享的社区.用户被显式链接，并且会有意无意向他们的朋友推荐商品或者服务。我们朋友购买过的商品更容易引起我们的注意力，并且会存储进情景记忆中，这些都会促进我们的消费。因此观察结果4就被很好的解释了。和朋友圈中流行的商品相比，用户对于那些不流行商品的消费更多是基于他们自己的喜好。因此，当物品被更多朋友消费时，评分会下降。



**模型和算法：**

在这一部分，我们将会展现我们的SPMF-MNAR模型，我们将从问题定义开始讲起。

问题定义:

在推荐系统中，我们有一个用户集合U(包括n个用户)和一个物品集合I(包括m个商品)，同时我们还有一个评分观察集合：



这里代表用户i对物品j的评分，集合O是有观察评分的用户和物品对构成的集合。实际生活中，通常是完整n\*m的评分矩阵R的一个子集。需要互译的是，用户对评分的观察过程就体现了用户的消费情况，因为用户在评分之前必须对该商品产生消费行为。因此，我们使用一个n∗m的消费矩阵X对R^0中的入口位置进行建模。对于每一个X , xij=1表示用户i已经消费并且评论过物品j,(rij∈R0可以被观察到)，xij=0表示评分不能被观察到（rij∉R0）。当涉及到社交推荐时，我们还有社交网络信息，可以表示用户之间的联系。构建向量，我们用它表示用户i的朋友们。其中g^i的第k个元素g^ik​代表用户i的第k个朋友，向量g^i的长度|g^i|代表了用户i的朋友数。社交推荐的任务，就是准确的预测用户的喜好或者评分以使得推荐符合用户的口味。

基于MNAR假设下的社交概率矩阵分解:

和现有社交推荐方法不同，我们要在社交推荐中引入MNAR假设，并且对评分的观察过程进行建模。如图1所示，一方面，我们根据社交影响和用户喜好生成了他们的消费矩阵。另一方面，采用传统的CF模型对用户消费过的商品做评分预测。

我们将概述SPMF-MNAR模型的生成过程，然后描述每一步的细节。对于每一次消费xij∈X和每一次评分值rij∈R，SPMF-MNAR模型将根据图三展示的模型过程生成。

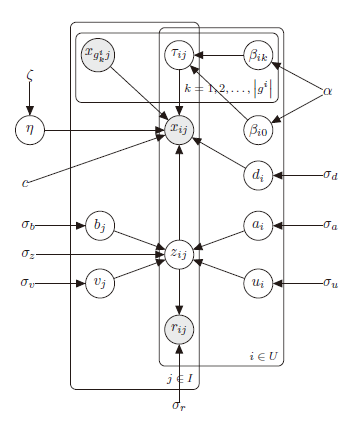


图3

步骤：

1）计算朋友对消费的影响：

2）对每个用户i而言，计算偏差：

计算接受阈值：

计算不同因素的重要性：

3）对每个物品j而言，计算偏差：

计算潜在价值：

4) 对于每个物品和用户的组合对(i,j):

计算用户对物品的喜好：

计算元素对消费的贡献：

当，说明消费受到了用户喜好的影响：；

当，说明消费是受到第k个朋友对用户的影响：

5) 对于每一个观察的评分rij :

根据用户的喜好推测用户评分：

在SPMF-MNAR模型中，我们做了这些基本假设：

假设一：

用户对物品的喜好可以通过概率矩阵分解生成【A. Mnih and R. R. Salakhutdinov, “Probabilistic matrix factorization,” in NIPS, 2008, pp. 1257–1264.】。对于每一个用户物品对，我们生成一个隐藏的连续变量zij,这个变量刻画了用户i对物品j的喜好。



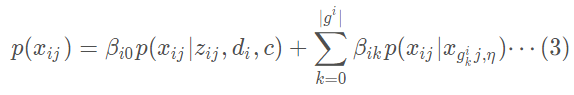
其中N()服从高斯分布，ui​代表着代表了用户i的D-维潜在喜好向量，vj​代表了物品j的D-维潜在价值。ai​和bj​代表了用户i和物品j的偏置。这里我们用ai​和bj​去刻画一种常见现象–有一些用户喜好很广泛，而一部分物品·质量很好。接着，我们可以基于用户喜好生成用户评分值如公式（2）所示：



这里，我们注意到我们简单的评分模型可以替代一些复杂的基于社交分析的推荐模型，比如SocialMF，TrustMF。但是我们同样承认，如果对社交影响建模可以进一步提升推荐效果。但是在这里我们只探究社交影响对用户评分过程（也就是购买行为）的影响，并探究它如何提升推荐效果。我们为了探究我们的基于社交模型的影响效果，所以我们排除了那些复杂的模型。

假设二：

从第三部分的分析可知，用户的消费时可以通过用户的喜好和社交影响对用户消费情况建模。对于每一个元素xij​∈X,有βi0​的概率购买行为是因为用户自己的喜好的概率 xij​∼p(xij​∣zij​,di​,c) ,有βik​的概率是被他第k个朋友影响 xij​∼p(xij​∣xgki​j,η​) .这里βik\beta\_{ik}βik​ 刻画了针对于每个人的社交影响的重要性。其中βI​=[βi0​,βi1​,…,βi∣gi∣​]，并且满足∑k=0∣gi∣​βik​=1，它服从Dirichlet分布，参数为α。基于这个生成过程，xij​可以展现为用户自己喜好和社交影响的随机组合。



为了更好的理解这些参数，我们引进多项式分布变量τij​∼Multi(βi​) 去描述每一个参量对购买xij​的贡献。τij​=[τij0​,τij1​,…,τij∣gi∣​] 是一个∣gi∣+1维的的二元随机变量，有∣gi∣+1种表达方式，每一种中只有一个元素为1，其余都为0。

假设三：

当xij​是通过用户个人喜好生成的时候，即τij0​=1时，伯努利对数模型可以用来描述用户趋向于购买他们喜欢的商品（就像分析结果1中显示的那样）：



其中σ(.)是逻辑函数。我们引入偏移量di​，di​定义了每个用户的接受阈值，去描述以下直觉认知：当用户对一个物品的喜好大于接受阈值时，即zij​>di​，用户就会倾向于购买某个物品，当用户对一个物品的喜好小于接受阈值时，即zij​<di​时，用户就会倾向于拒绝购买某个物品。同时，我们引进了一个比例参数c，c > 0 去控制用户喜好性zij​对消费行为xij​的影响。当比例参数c值变大时，用户会对zij​更敏感。换句话说， zij​的轻微扰动都会带来xij​的大的变化。总结一下,我们可以发现zij​越大， p(xij​=1) 就会越大，也就意味着用户更有可能去购买这个商品，和我们在观察结果1是吻合的。

假设四：

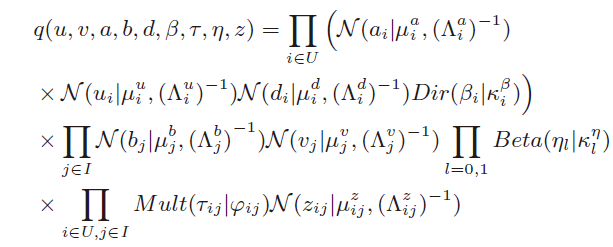
当用户的购买行为xij​是他第k个朋友影响生成的时候，τijk​=1，我们对  
xij​进行伯努利分布和他第k个朋友的购买情况xgki​j​的伯努利参数进行建模。



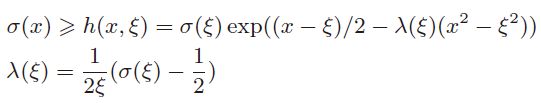
参数η=[η0​,η1​]代表了用户消费行为对他指定朋友消费行为的条件依赖性。η1​=p(xij​=1∣xkj​=1,τijk​=1)意味着用户会跟随他朋友的消费行为消费的概率。η0​=p(xij​=1∣xkj​=0,τijk​=1)意味着当用户的朋友没有消费这个商品时，用户购买的概率。事实上，用户倾向于购买他们朋友购买过的商品，所以参数η0​和参数η1​应该有显著的不同。我们在第五部分实验结果也验证了这个观点，η1​的后验概率要远大η0​。

**推理猜想：**

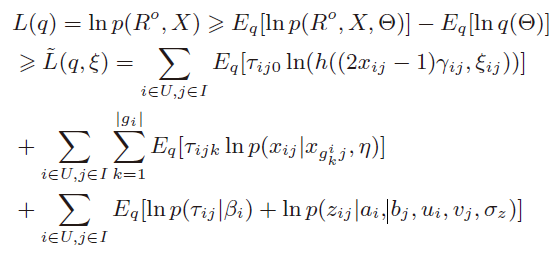
考虑到SPMF-MNAR模型中的后验概率很难精确计算，所以 我们建立了一种有效的估计方法去根据变分推断去计算后验概率，并且平均场理论要求我们把这些分离的隐藏变量划分成不相交的组，并且这些变量都有它们自己的变量参数管控。此外，我们可以它们的相关条件指出这些变量的分布情况【A. Honkela, M. Tornio, T. Raiko, and J. Karhunen, “Natural conjugate gradient in variational inference,” in International Conference on Neural Information Processing. Springer, 2007, pp. 305–314.】。也就是说，我们可以定义如下的变量分布：



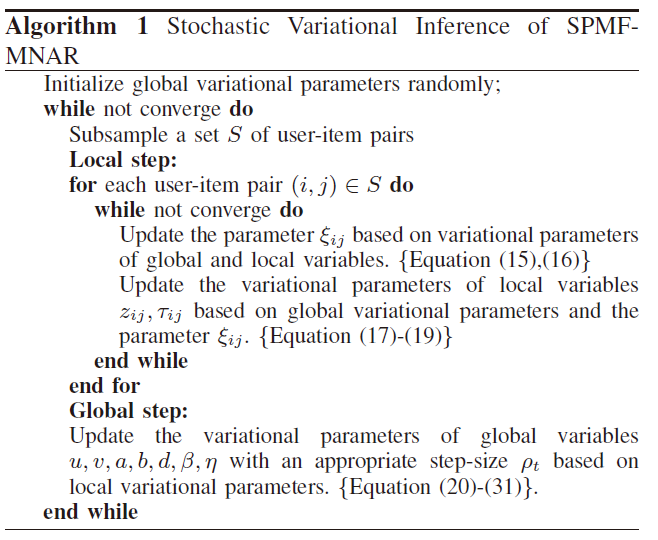
需要注意，伯努利对数似然函数不承认共轭先验。为了解决这个问题，我们应用高斯函数降低这个逻辑函数的约束性。



通过增加的参数ε我们可以获得一个更低边界值的逻辑函数。因此我们有了以下目标函数，对观察值的似然函数取log。



其中Θ≡{u,v,a,b,d,β,τ,η,z}代表潜变量，λij​=c(zij​−di​).我们使用坐标上升法通过不变降低下限L(q,ε)交替优化变量参数,潜在局部变量zij​,τij​,εij​都有n×m个入口，如果在每一轮迭代中去优化每一个入口是非常低效的，事实上，我们更希望去学习一些全局变量的参数，比如u,v,a,b,d,β,η.随机变量推断法(SVI)可以快速优化全局变量参数。我们可以通过每次只使用子集数据就能学习到全局变量参数。在algorithm1中，我们展示了我们随机变量推断算法的为代码，细节在附录中。



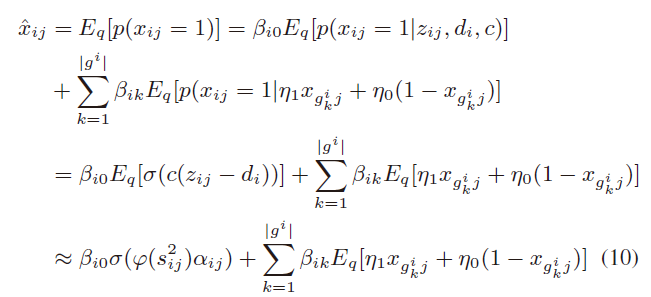
**复杂度分析**：时间度主要是更新变量参数导致的，在每一个局部过程中，我们只更新针对在集合S中用户物品对(i,j)的局部参数，时间复杂度为O(tl​(∣S∣+∣G∣)),∣S∣代表了集合S中用户物品对的数量，∣G∣表示数据集中变得数量，tl​代表了局部过程中的迭代次数.在全局过程中，我们只是将集合S中的参数求和，所以更新全局变量的计算复杂度为O((n+m)D^3+|S|+|G|) ,其中n和m代表数据集中用户和物品的数量，D代表潜在向量ui​,vj​的维度。因此一次的时间复杂度为O((n+m)D^3+tl(|S|+|G|)).事实上，我们会让D和tl为一个固定的小数字，用户物品对|S|为评分数量的两倍。因此我们的算法在稀疏推荐矩阵上是高效的。

**预测：**

一旦生成了合适的后验概率，我们就能预测用户i对物品j的评分。预测评分的计算公式如下式所示：



这里我们可以增加一个新步骤去计算zij​的后验值。需要注意的是在我们的SPMF-MNAR模型中是采用MNAR假设的，所以我们可以根据预测的购买情况作推荐，预测的购买情况是融合了用户喜好和社交影响的。



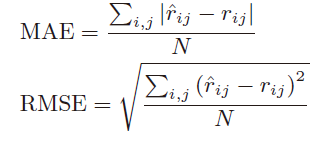
我们使用重标逻辑函数去近似逻辑函数使它和逻辑函数σ(.)有相同的切线斜率，并且ϕ(x)=(1+πx/8)(−1/2),sij2​=Dq​[c(zij​−di​)],αij​=Eq​[c(zij​−di​)]。

**实验和结果：**

在这一部分，我们在真实数据集上检验了我们的SPMF-MNAR模型，首先我们介绍一下我们的实验方案。

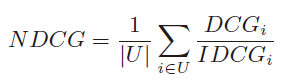
实验方案：我们将用第三部分介绍到的两个数据集Epinions和Ciao进行我们的实验。数据集包括了用户的消费情况，评分情况和社交关系。使用五折交叉验证的方式进行训练和测试。并且如果xij​=1在测试集中出现了，我们就会让训练集中xij​=0，去表示rij​没有被观察到。最优实验参数的获得方式来源有两个，前人推荐的最优参数和网格搜索法。对于一些超参数，我们在两个数据集中都设置σr2​=0.1,σd2​=0.2,D=10,其他为1，除了c有所不同，如图5所示。步长的设置我们参考了论文【M. D. Hoffman, D. M. Blei, C. Wang, and J. Paisley, “Stochastic variational inference,” The Journal of Machine Learning Research, vol. 14, no. 1, pp. 1303–1347, 2013.】中设置ρt​=(t0​+t)−ϑ,t0​=0,ϑ=0.8,对于每一次采样的集合S，我们选择所有xij​=1的数据，并且随机选择等数量的xij​=0的数据去训练我们的模型。

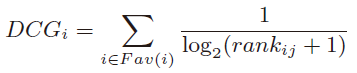
评估指标：我们采用下述评估指标去检验预测准确率和推荐效果。MAE平均绝对误差，RMSE均方误差：



其中N是测试集中评分数量，rij​是用户i对物品j评分的实际观察值，r^ij​是用户i对物品j的预测值。

NDCG(评价排序质量)：NDCG指标被广泛应用于信息抽取中，评估按照重要性降序排列的排序质量。在推荐中，NDCG的计算方法如下所示：



DCGi​的定义如下面的公式所示， IDCGi​是DCGi​最优排序的理想值。

其中rank ij ​代表了物品j在用户i的推荐列表上的排序位置; Fav(i)代表了用户i在测试集中喜欢的物品的集合。正如别人近期的工作中所提到的，在测试集中高评分的物品就可以认为是喜欢的物品（比如5星中3星的物品），NDCG可以看作一种寻找所有喜欢的物品的难易程度，NDCG值越大，就意味着喜欢的物品在列表中越靠前的位置。

对比方法：为了证明SPMF-MNAR模型的有效性，我们和以下方法进行对比:

**基线:** Useravg,PMF、Useravg是一个非社交关系的简单基线模型，通过用于的平均分取预测缺失的分数。PMF是一个非社交的评分数据的经典概率矩阵分解模型。

**不采用MNAR假设的社交推荐算法:**SoRec，RSTE，SocialMF，TrustMF，TrustSVD都是社交推荐中最优的模型。需要注意的是，这些社交推荐模型都没有采用MNAR假设。  
 **采用MNAR假设的非社交推荐算法:**GPMF是一种采用MNAR假设的非社交推荐算法最优模型。我们也自己设计了一个SPMF-MNAR-nos模型作为比较去验证对社交关系建模对评分观察过程的影响。SPMF-MNAR-nos是SPMF-MNAR的简化版，排除了社交影响，只通过用户喜好生成用户购买情况。

**实验结果和分析：**

表2展现了这是个推荐算法在两个数据集上的MAE和RMSE。我们发现SPMF-MNAR模型在所有数据集上都优于其他模型。SPMF-MNAR模型比其它方法的改善主要体现在两个方面，

（1）在社交推荐中的评分不是随机丢失的，如果方法采用MAR的假设，忽略用户是否消费这个物品和用户评分的联系，预测结果会受损。

（2）我们在推荐系统中融入社交信息。

和以往探究社交影响对用户评分的研究不同，SPMF-MNAR模型致力于探究社交影响对用户评分过程也就是是否会产生购买行为进行研究，并且观察它如何改善推荐效果。事实上，我们在数据集上的分析表明在内容共享的社交网络中，相互连接的用户通常会不自觉的相互推荐，因此他们的消费会被社交关系强烈影响。所以，SPMF-MNAR的推荐效果比不应用社交关系的SPMF-MNAR-nos模型更好。

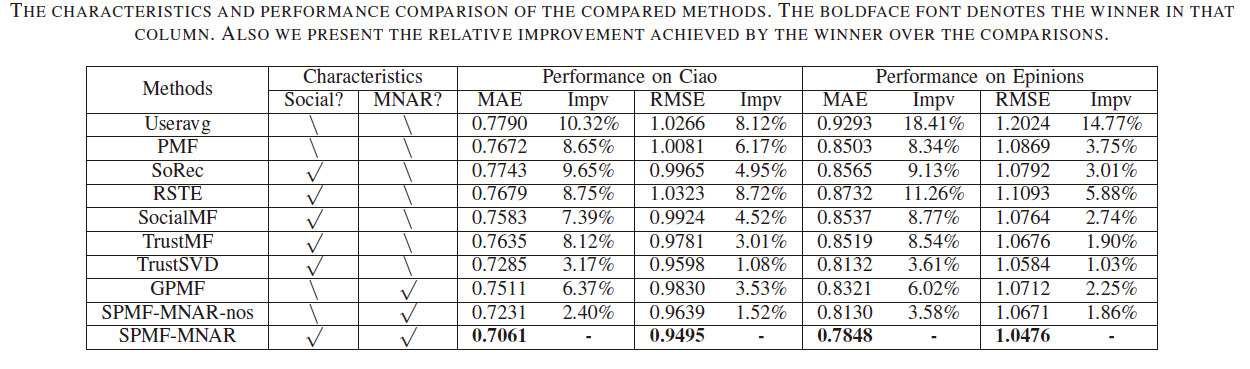


表2

NDCG比较结果如下，下图4展现了不同方法的NDCG表现情况，NDCG刻画了排序质量，NDCG值越大，效果越好。我们观察到，SPMF-MNAR再一次比其它方法效果好。尤其，在NDCG上SPMF-MNAR比有了显著的提升，在数据集Ciao上提升了22.3%，数据集Epinions上提升了19.3%。这个结果充分展现了社交关系对用户消费情况的影响。

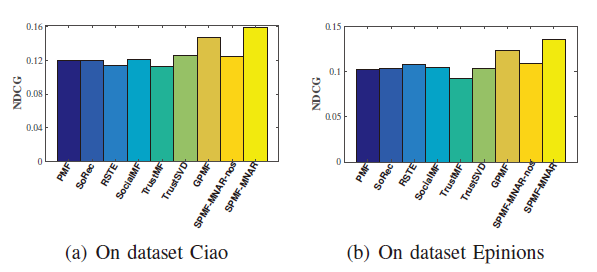


图4

我们还做了一个实验去探究参数c对SPMF-MNAR模型的影响，其中c是一个规模参数控制用户喜好zij​对他消费情况xij​的影响。图5展现了不同c值下的MAE，c越大，MAE一开始变小，推荐效果变好，这是因为在社交推荐系统中，用户消费情况不是完全和用户喜好无关的。用户的消费也反映了用户对物品的喜好。因此，当我们增加用户喜好对用户消费情况的影响时，反过来我们也可以从用户的消费情况推断出更多的用户喜好。但是，当c大于判决阈值0.5时，增加c推荐效果变差。因为c越大，我们使用更依赖于消费信息去推断用户的喜好，其它重要信息包括用户评分被忽略了。因此，只有c设置为一个合适的值，比如c=0.5，它才能平衡这些信息的重要性，SPMF-MNAR模型才能达到最佳效果。

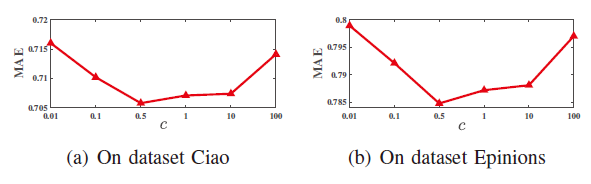


图5

我们在表3中展现了η0​和η1​的最优参数分布情况。η0​和η1​表现了用户消费和他们朋友之间消费的依赖性。η1​=p(xij​=1∣xkj​=1,τijk​=1)意味着用户会跟随他指定朋友消费情况消费的概率。η0​=p(xij​=1∣xkj​=0,τijk​=1)意味着用户的第k个朋友没有消费该商品他消费的概率。从表3我们得知，η1​的数学期望大于0.99，η0​的数学期望小于0.01。这个结果表明了社交服从效应，用户会根据朋友的购买情况决定是否购买商品。所以我们在对用户购买情况建模时，我们需要考虑社交影响的影响可能会让我们更好的推断出用户真实的喜好。

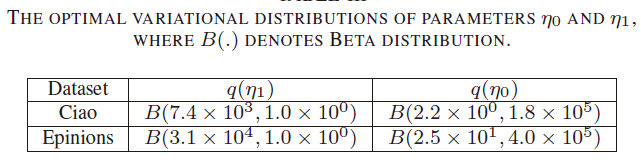


表3

去探究用户的主观能动性和用户的接受阈值d的联系是有意义的。图6展现了不同用户主观能动性不同时判决阈值d的后验期望值。用户的接受阈值d比较大时，用户会对物品更挑剔，用户通常不会接受朋友推荐的东西。与之相反的是，判决阈值d小的用户通常比较宽容，他们通常会消费更多的东西。

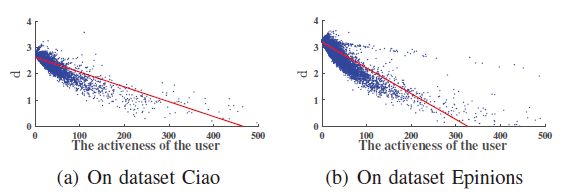


图6

**结论：**

本文中，我们对社交推荐的真实数据集自习分析，并且发现评分过程依赖于用户的喜好和社交关系。因此我们引入MNAR假设进入社交推荐系统，并且提出一种新的推荐方法SPMF-MNAR 通过一种概率生成模型。和现有推荐系统研究社交影响对用户评分的预测不同，我们把关注点放在了探究社交关系对评分过程的影响，也就是是否会购买。我们的实验结果展现了模型的有效性。

将来我们会探究更多会影响用户消费的因素，比如物品流行性和用户位置，我们将扩展SPMF-MNAR模型以获得更好的效果。