基于 Logistic 函数的贝叶斯概率矩阵分解算法

方耀宁* 郭云飞 兰巨龙
(国家数字交换系统工程技术研究中心 郑州 450002)

摘 要：在协同过滤推荐系统中，矩阵分解是一种非常有效的工具。贝叶斯概率矩阵分解模型具有预测精度高的优点，但不能表示潜在因子之间的非线性关系。针对该问题，该文提出一种基于 Logistic 函数的改进贝叶斯概率矩阵分解模型，并使用马尔科夫链蒙特卡罗方法进行训练。在两组真实数据集合上的实验表明，基于 Logistic 函数的贝叶斯概率矩阵分解算法能够明显提高预测准确性，有效缓解数据稀疏性问题。

关键词：推荐系统；信息处理；协同过滤；贝叶斯概率矩阵分解；Logistic 函数
中图分类号：TP393 **文献标识码**：A **文章编号**：1009-5896(2014)03-0715-06
DOI: 10.3724/SP.J.1146.2013.00534

A Bayesian Probabilistic Matrix Factorization Algorithm Based on Logistic Function

Fang Yao-ning Guo Yun-fei Lan Ju-long
(*National Digital Switching System Engineering and Technological R&D Center*, *Zhengzhou* 450002, *China*)

Abstract: The matrix factorization is one of the most powerful tools in collaborative filtering recommender systems. The Bayesian Probabilistic Matrix Factorization (BPMF) model has advantages of high prediction accuracy, but can not capture non-linear relationships between latent factors. To address this problem, an improved model is proposed based on the Logistic function and Markov Chain Monte Carlo is used to train the proposed model. Experiments on two real-world benchmark datasets show significant improvements in prediction accuracy compared with several state-of-the-art methods for recommendation tasks.

Key words: Recommender system; Information processing; Collaborative filtering; Bayesian Probabilistic Matrix Factorization (BPMF); Logistic function

1 引言

当前，推荐系统(Recommender Systems, RS)是电子商务领域的研究热点之一。与基于内容的(Content-based)推荐算法相比，协同过滤(Collaborative Filtering, CF)推荐算法无须对内容进行解析，只利用历史行为信息便可以进行个性化推荐，在过去十年中取得了丰富的研究成果。其中，基于矩阵分解(Matrix Factorization, MF)的潜在因子模型(latent factor model)在预测准确性和稳定性上得到了最为广泛的认可[1]。

推荐算法中，通常假设评分矩阵 $\boldsymbol{R}\in\mathbb{R}^{N\times M}$ 是低秩矩阵，可以用两个低秩矩阵的乘积近似表示，即 $\boldsymbol{R}\approx\boldsymbol{U}^{\mathrm{T}}\boldsymbol{V}$，其中 $\boldsymbol{U}\in\mathbb{R}^{d\times N}$，$\boldsymbol{V}\in\mathbb{R}^{M\times d}$，$d\ll\min(N,M)$。评分矩阵 $\boldsymbol{R}$ 中未知的数据 $\boldsymbol{R}_{ij}$，由 $\boldsymbol{U}_i^{\mathrm{T}}\boldsymbol{V}_j$ 来进行预测。文献[2]首先把奇异值分解(Singular Value Decomposition, SVD)应用到推荐系统领域，文献[3]通过增加用户和项目的偏置(bias)对奇异值分解算法进行了改进。Koren[4]提出了利用隐含反馈信息的 SVD++模型，取得了较高的预测准确性，但计算量较大。非负矩阵分解算法是把评分矩阵分解为两个非负矩阵的乘积，但在推荐算法中预测精度并不高[5]。概率矩阵分解(Probabilistic Matrix Factorization, PMF)模型提供了对 SVD 模型进行正则化(regularization)的概率解释[6]，文献[7]进一步将概率矩阵分解模型扩展到贝叶斯概率矩阵分解(Bayesian Probabilistic Matrix Factorization, BPMF)模型，并使用马尔科夫链蒙特卡罗(Markov Chain Monte Carlo, MCMC)方法进行训练。Bayesian PMF 算法的预测准确性比较高，而且不需要设定学习率和正则化系数，但仍是一种线性模型，不能表示潜在因子之间的非线性关系。

核矩阵分解算法(kernelized matrix factorization)虽然能够表征潜在因子间的非线性联系，但是计算量较大并不实用[8,9]。此外，项目间的隐含关系，用户的社交信息，上下文也可以用来提高矩阵分解算法的性能[10,11]。最近，Mackey 等人[12]提出“分而

2013-04-19 收到，2013-07-29 改回
国家 973 计划项目(2012CB315901)和国家 863 计划项目(2011AA01A103)资助课题
*通信作者：方耀宁 fyn07@163.com

治之”的思想，把原始矩阵拆分成多个小矩阵分别进行分解。与之类似，文献[13]假设原始矩阵是局部低秩的，把原始矩阵分解为多个低秩矩阵，然后再进行分解。文献[14]和文献[15]在 Netflix 和 MovieLens 数据集合上对主流的推荐算法进行了较为详细的对比分析。

本文利用 Logistic 函数来表征潜在因子间的非线性关系，在不增加计算复杂度的前提下，建立了 L-BPMF(Logistic Bayesian Probabilistic Matrix Factorization)模型，并使用马尔科夫链蒙特卡罗方法训练 L-BPMF 模型。在两种真实数据集合上的实验结果表明，L-BPMF 比主流推荐算法的预测准确性都要好，能够明显缓解数据稀疏性问题[15]。

2 贝叶斯概率矩阵分解

概率矩阵分解(Probabilistic Matrix Factorization, PMF)假设用户和项目的特征向量矩阵 $\boldsymbol{M}$, $\boldsymbol{N}$ 都服从高斯分布，不同用户、项目的概率分布相互独立[6]。

$$\left.\begin{aligned} p(\boldsymbol{U}\mid\boldsymbol{\mu}_{\boldsymbol{U}},\boldsymbol{\Lambda}_{\boldsymbol{U}})&=\prod_{i=1}^{N}\mathcal{N}\left(\boldsymbol{U}_i\mid\boldsymbol{\mu}_{\boldsymbol{U}},\boldsymbol{\Lambda}_{\boldsymbol{U}}^{-1}\right)\\ p(\boldsymbol{V}\mid\boldsymbol{\mu}_{\boldsymbol{V}},\boldsymbol{\Lambda}_{\boldsymbol{V}})&=\prod_{j=1}^{M}\mathcal{N}\left(\boldsymbol{V}_j\mid\boldsymbol{\mu}_{\boldsymbol{V}},\boldsymbol{\Lambda}_{\boldsymbol{V}}^{-1}\right)\end{aligned}\right\}\tag{1}$$

进而，把用户对项目的评分变成一个概率问题：

$$p(\boldsymbol{R}\mid\boldsymbol{U},\boldsymbol{V},\alpha)=\prod_{i=1}^{N}\prod_{j=1}^{M}\left[\mathcal{N}\left(\boldsymbol{R}_{ij}\mid\boldsymbol{U}_i^{\mathrm{T}}\boldsymbol{V}_j,\alpha^{-1}\right)\right]^{I_{ij}}\tag{2}$$

其中 $\mathcal{N}(x\mid\boldsymbol{\mu},\alpha^{-1})$ 是期望为 $\boldsymbol{\mu}$，方差为 α^{-1} 的高斯分布。I_{ij} 是示性函数，如果用户 i 选择了项目 j，那么 $I_{ij}=1$，否则 $I_{ij}=0$。

参数 $\{\boldsymbol{\mu}_{\boldsymbol{V}},\boldsymbol{\mu}_{\boldsymbol{U}}\}$ 一般都可以设定为 0，但参数 $\{\boldsymbol{\Lambda}_{\boldsymbol{U}},\boldsymbol{\Lambda}_{\boldsymbol{V}}\}$ 的选择对于算法的预测性能有着重要影响，寻找合适参数往往是一件费时耗力的事情[6,7]。如图 1 所示，Bayesian PMF 模型进一步设定 $\boldsymbol{\Theta}_{\boldsymbol{U}}=\{\boldsymbol{\mu}_{\boldsymbol{U}},\boldsymbol{\Lambda}_{\boldsymbol{U}}\}$, $\boldsymbol{\Theta}_{\boldsymbol{V}}=\{\boldsymbol{\mu}_{\boldsymbol{V}},\boldsymbol{\Lambda}_{\boldsymbol{V}}\}$ 的先验分布为高斯-威沙特分布(Gaussian-Wishart distribution)，把参数 $\{\boldsymbol{\Lambda}_{\boldsymbol{U}},\boldsymbol{\Lambda}_{\boldsymbol{V}}\}$ 整合到模型内部，避免了寻找最优参数的过程[7]。

$$\left.\begin{aligned} p(\boldsymbol{\Theta}_{\boldsymbol{U}}\mid\boldsymbol{\Theta}_0)&=p\left(\boldsymbol{\mu}_{\boldsymbol{U}}\left|\boldsymbol{\Lambda}_{\boldsymbol{U}},\boldsymbol{\Theta}_0\right.\right)p\left(\boldsymbol{\Lambda}_{\boldsymbol{U}}\left|\boldsymbol{\Theta}_0\right.\right)\\ &=\mathcal{N}\left(\boldsymbol{\mu}_{\boldsymbol{U}}\left|\boldsymbol{\mu}_0,(\beta_0\boldsymbol{\Lambda}_{\boldsymbol{U}})^{-1}\right.\right)\mathcal{W}\left(\boldsymbol{\Lambda}_{\boldsymbol{U}}\left|\boldsymbol{W}_0,\boldsymbol{\nu}_0\right.\right)\\ p(\boldsymbol{\Theta}_{\boldsymbol{V}}\mid\boldsymbol{\Theta}_0)&=p(\boldsymbol{\mu}_{\boldsymbol{V}}\mid\boldsymbol{\Lambda}_V,\boldsymbol{\Theta}_0)p(\boldsymbol{\Lambda}_{\boldsymbol{V}}\mid\boldsymbol{\Theta}_0)\\ &=\mathcal{N}\left(\boldsymbol{\mu}_{\boldsymbol{V}}\left|\boldsymbol{\mu}_0,(\beta_0\boldsymbol{\Lambda}_{\boldsymbol{V}})^{-1}\right.\right)\mathcal{W}\left(\boldsymbol{\Lambda}_V\left|\boldsymbol{W}_0,\boldsymbol{\nu}_0\right.\right)\end{aligned}\right\}\tag{3}$$

其中 $\boldsymbol{\Theta}_0=\{\boldsymbol{\mu}_0,\boldsymbol{\nu}_0,\boldsymbol{W}_0,\alpha,\beta_0\}$，$\mathcal{W}(\boldsymbol{\Lambda}\mid\boldsymbol{W}_0,\boldsymbol{\nu}_0)$ 是自由度为 $\boldsymbol{\nu}_0$，协方差矩阵为 $\boldsymbol{W}_0$ 的威沙特分布。

BPMF 模型一般用马尔科夫链蒙特卡罗方法进行训练，虽然预测误差较小，但仍然是一种线性模型。线性模型不能反映用户特征 $\boldsymbol{U}$ 和项目特征 $\boldsymbol{V}$ 之间的非线性关系，这一定程度上限制 BPMF 的性能，特别是在数据稀疏条件下提取潜在信息的能力。

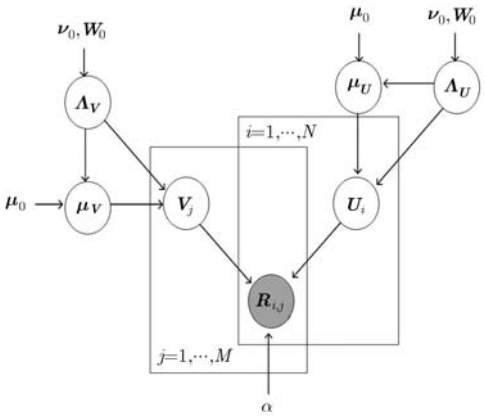

图1 Bayesian PMF模型[7]

3 基于 Logistic 函数的 BPMF 算法

3.1 Logistic BPMF 模型

Logistic 函数是机器学习领域中常用的一种 S 型函数，定义域为 $(-\infty,+\infty)$，值域为(0,1)。Logistic 函数在定义域内单调连续，呈现出先缓慢增长，然后加速增长，最后逐渐稳定的趋势，能够较好反映生物种群发展、神经元非线性感知、人类认知学习过程等[16]。受数理情感学中情感强度第一定律启发，本文用 Logistic 函数的横轴表示用户感知到的“刺激”，纵轴表示用户评分：用户受到的“刺激”为正，则评分大于评分均值；“刺激”越强烈，评分越高；原点附近评分随“刺激”变化较快，属于敏感区域；偏离原点处评分随“刺激”变化缓慢，属于麻木区域[16,17]。

Logistic BPMF 模型如图 2 所示，其核心改进是假设用户 i 对项目 j 的评分 $\boldsymbol{R}_{ij}$ 服从均值为 $\boldsymbol{B}_i\, g(\boldsymbol{U}_i^{\mathrm{T}}\boldsymbol{V}_j)$，方差为 α^{-1} 的高斯分布，即

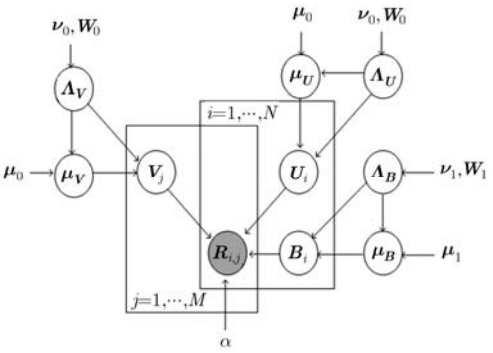

图2 Logistic BPMF模型

$$p\left(\boldsymbol{R}\middle|\boldsymbol{U},\boldsymbol{V},\boldsymbol{B},\alpha^{-1}\right)=\prod_{i=1}^{N}\prod_{j=1}^{M}\left[\boldsymbol{\mathcal{N}}\left(\boldsymbol{R}_{ij}\middle|\boldsymbol{B}_i\, g\left(\boldsymbol{U}_i^{\mathrm{T}}\boldsymbol{V}_j\right),\alpha^{-1}\right)\right]^{I_{ij}} \tag{4}$$

其中 $g(x)$表示 Logistic 函数， $\boldsymbol{B}_i$是表示用户 i 评分尺度的参数。不失一般性，同样假设 $\boldsymbol{B}$ 服从均值为 $\boldsymbol{\mu}_B$，方差为 $\boldsymbol{\Lambda}_{\boldsymbol{B}}^{-1}$ 的高斯分布。与 BPMF 模型相似，为了便于 MCMC 训练过程中后验概率的计算，设定超参数 $\boldsymbol{\Theta}_{\boldsymbol{B}}=\{\boldsymbol{\mu}_{\boldsymbol{B}},\boldsymbol{\Lambda}_{\boldsymbol{B}}\}$ 的先验分布为高斯-威沙特分布[7]。

$$\begin{aligned}p\left(\boldsymbol{\Theta}_{\boldsymbol{B}}\middle|\boldsymbol{\Theta}_0\right)&=p\left(\boldsymbol{\mu}_{\boldsymbol{B}}\middle|\boldsymbol{\Lambda}_{\boldsymbol{B}},\boldsymbol{\Theta}_0\right)p\left(\boldsymbol{\Lambda}_{\boldsymbol{B}}\middle|\boldsymbol{\Theta}_0\right)\\&=\mathcal{N}\left(\boldsymbol{\mu}_{\boldsymbol{B}}\middle|\boldsymbol{\mu}_1,\left(\beta_0\boldsymbol{\Lambda}_{\boldsymbol{B}}\right)^{-1}\right)\varpi\left(\boldsymbol{\Lambda}_{\boldsymbol{B}}\middle|\boldsymbol{W}_1,\nu_1\right)\end{aligned} \tag{5}$$

此时，$\boldsymbol{\Theta}_0$ 为 $\{\boldsymbol{\mu}_0,\boldsymbol{\mu}_1,\boldsymbol{\nu}_0,\boldsymbol{\nu}_1,\boldsymbol{w}_0,\boldsymbol{w}_1,\alpha,\beta_0\}$。虽然贝叶斯模型(BPMF, L-BPMF)的初始化参数比概率矩阵分解模型多，但是贝叶斯模型对初始化参数的取值并不敏感。一般设定：$\boldsymbol{w}_0$ 和 $\boldsymbol{w}_1$ 为单位矩阵，$\alpha=2$，$\beta_0=2$，$\boldsymbol{\mu}_0=0$，$\boldsymbol{\mu}_{\boldsymbol{B}}=2m$，$\boldsymbol{\nu}_0=\mathrm{rank}(\boldsymbol{w}_0)$，$\boldsymbol{\nu}_B=\mathrm{rank}(\boldsymbol{w}_1)$。其中，$m$ 为系统中用户评分均值。

用户 i 对项目 j 评分 $\boldsymbol{R}_{ij}^{*}$ 的概率分布由式(6)确定。

$$\begin{aligned}&p\left(\boldsymbol{R}_{ij}^{*}\middle|\boldsymbol{R},\boldsymbol{\Theta}_0\right)\\&=\iint p\left(\boldsymbol{R}_{ij}^{*}\middle|\boldsymbol{U}_i,\boldsymbol{V}_j,\boldsymbol{B}_i\right)p\left(\boldsymbol{U},\boldsymbol{V},\boldsymbol{B}\middle|\boldsymbol{R},\boldsymbol{\Theta}_{\boldsymbol{U}},\boldsymbol{\Theta}_{\boldsymbol{V}},\boldsymbol{\Theta}_{\boldsymbol{B}}\right)\\&\quad\cdot p\left(\boldsymbol{\Theta}_{\boldsymbol{U}},\boldsymbol{\Theta}_{\boldsymbol{V}},\boldsymbol{\Theta}_{\boldsymbol{B}}\middle|\boldsymbol{\Theta}_0\right)\mathrm{d}\{\boldsymbol{U},\boldsymbol{V},\boldsymbol{B}\}\mathrm{d}\{\boldsymbol{\Theta}_{\boldsymbol{U}},\boldsymbol{\Theta}_{\boldsymbol{V}},\boldsymbol{\Theta}_{\boldsymbol{B}}\}\end{aligned} \tag{6}$$

一般来说，式(6)中条件联合概率 $p(\boldsymbol{U},\boldsymbol{V},\boldsymbol{B}\mid\boldsymbol{R},\boldsymbol{\Theta}_{\boldsymbol{U}},\boldsymbol{\Theta}_{\boldsymbol{V}},\boldsymbol{\Theta}_{\boldsymbol{B}})$ 不容易得到，导致整个二重积分无法直接进行计算。一种求解的思路是利用马尔科夫链蒙特卡罗方法对$\{\boldsymbol{U},\boldsymbol{V},\boldsymbol{B}\}$进行抽样，然后根据式(7)来近似计算。

$$p\left(\boldsymbol{R}_{ij}^{*}\mid\boldsymbol{R},\boldsymbol{\Theta}_0\right)\approx\frac{1}{T}\sum_{t=1}^{\mathrm{T}}p\left(\boldsymbol{R}_{ij}^{*}\middle|\boldsymbol{U}_i^{t},\boldsymbol{V}_j^{t},\boldsymbol{B}_i^{t}\right) \tag{7}$$

3.2 贝叶斯推断与吉布斯抽样

. 贝叶斯推断(Bayesian inference)是将先验的思想和样本数据相结合，得到后验分布，然后根据后验分布进行统计推断。贝叶斯推断的精度受到样本数量和先验分布准确性的影响。在先验分布一定的情况下，样本数量越大则推断精度越高。本文在训练 Logistic BPMF 的过程中，使用吉布斯抽样进行贝叶斯推断。吉布斯抽样(Gibbs sampling)是一种典型的 MCMC 方法，适用于联合概率未知，但条件概率容易获取的情况。首先利用条件概率构造平稳分布为所求联合概率的马尔科夫链，然后进行 T 次抽样，此时的样本$\{\boldsymbol{U},\boldsymbol{V},\boldsymbol{B}\}$可以近似认为是来自联合概率 $p(\boldsymbol{U},\boldsymbol{V},\boldsymbol{B}\mid\boldsymbol{R},\boldsymbol{\Theta}_{\boldsymbol{U}},\boldsymbol{\Theta}_{\boldsymbol{V}},\boldsymbol{\Theta}_{\boldsymbol{B}})$ 的抽样，最后利用式(7)进行评分预测。

使用吉布斯抽样方法进行贝叶斯推断，条件后验概率必须要有显示解，即明确的新样本产生规则。L-BPMF 模型中，在已知其它参数的条件下，$\boldsymbol{U}_i$ 的条件后验概率为

$$\begin{aligned}&p\left(\boldsymbol{U}_i\mid\boldsymbol{R},\boldsymbol{V},\boldsymbol{B},\boldsymbol{\Theta}_{\boldsymbol{U}},\alpha\right)\\&\propto\prod_{j=1}^{M}\left[\boldsymbol{\mathcal{N}}\left(\boldsymbol{R}_{ij}\middle|\boldsymbol{B}_i g\left(\boldsymbol{U}_i^{\mathrm{T}}\boldsymbol{V}_j\right),\alpha^{-1}\right)\right]^{I_{ij}}p\left(\boldsymbol{U}_i\middle|\boldsymbol{\mu}_{\boldsymbol{U}},\boldsymbol{\Lambda}_{\boldsymbol{U}}\right)\end{aligned} \tag{8}$$

为了简化式(9)的形式，本文对 Logistic 函数进行麦克劳林展开，于是

$$\begin{aligned}&\boldsymbol{\mathcal{N}}\left(\boldsymbol{R}_{ij}\middle|\boldsymbol{B}_i g\left(\boldsymbol{U}_i^{\mathrm{T}}\boldsymbol{V}_j\right),\alpha^{-1}\right)\\&\approx\boldsymbol{\mathcal{N}}\left(\boldsymbol{R}_{ij}\middle|\boldsymbol{B}_i\left(\frac{1}{2}+\frac{\boldsymbol{U}_i^{\mathrm{T}}\boldsymbol{V}_j}{4}\right),\alpha^{-1}\right)\\&=\boldsymbol{\mathcal{N}}\left(\boldsymbol{R}_{ij}\middle|\boldsymbol{B}_i\left(\frac{1}{2}+\frac{\boldsymbol{U}_i^{\mathrm{T}}\boldsymbol{V}_j}{4}\right),\alpha^{-1}\right)\\&=\boldsymbol{\mathcal{N}}\left(\frac{4\boldsymbol{R}_{ij}-2\boldsymbol{B}_i}{\boldsymbol{B}_i}\middle|\boldsymbol{U}_i^{\mathrm{T}}\boldsymbol{V}_j,\left(\alpha\frac{\boldsymbol{B}_i^2}{16}\right)^{-1}\right)\end{aligned} \tag{9}$$

根据共轭先验分布理论：若方差已知，高斯分布均值的共轭先验分布是高斯分布[7]，利用配凑平方和的方法可以得到

$$\begin{aligned}&p\left(\boldsymbol{U}_i\mid\boldsymbol{R},\boldsymbol{V},\boldsymbol{B},\boldsymbol{\Theta}_{\boldsymbol{U}},\alpha\right)\\&\propto p\left(\boldsymbol{R}\middle|\boldsymbol{U}_i,\boldsymbol{V},\boldsymbol{B},\alpha\right)p\left(\boldsymbol{U}_i\middle|\boldsymbol{\Theta}_{\boldsymbol{U}}\right)\\&\approx\prod_{j=1}^{M}\left[\boldsymbol{\mathcal{N}}\left(\frac{4\boldsymbol{R}_{ij}-2\boldsymbol{B}_i}{\boldsymbol{B}_i}\middle|\boldsymbol{U}_i^{\mathrm{T}}\boldsymbol{V}_j,\left(\alpha\frac{\boldsymbol{B}_i^2}{16}\right)^{-1}\right)\right]^{I_{ij}}\\&\cdot\boldsymbol{\mathcal{N}}\left(\boldsymbol{U}_i\middle|\boldsymbol{\mu}_{\boldsymbol{U}},\boldsymbol{\Lambda}_{\boldsymbol{U}}\right)\\&\propto\boldsymbol{\mathcal{N}}\left(\boldsymbol{U}_i\middle|\boldsymbol{\mu}_{Ui}^{*},\left[\boldsymbol{\Lambda}_{Ui}^{*}\right]^{-1}\right)\end{aligned} \tag{10}$$

其中 $\boldsymbol{\Lambda}_{Ui}^{*}=\boldsymbol{\Lambda}_{\boldsymbol{U}}+\alpha\frac{\boldsymbol{B}_i^2}{16}\sum_{j=1}^{M}\left[\boldsymbol{V}_j\boldsymbol{V}_j^{\mathrm{T}}\right]^{I_{ij}}$，$\boldsymbol{\mu}_{Ui}^{*}=\left[\boldsymbol{\Lambda}_{Ui}^{*}\right]^{-1}\cdot\left(\alpha\frac{\boldsymbol{B}_i^2}{16}\sum_{j=1}^{M}\left[\boldsymbol{V}_j\frac{4\boldsymbol{R}_{ij}-2\boldsymbol{B}_i}{\boldsymbol{B}_i}\right]^{I_{ij}}+\boldsymbol{\Lambda}_{\boldsymbol{U}}\boldsymbol{\mu}_{\boldsymbol{U}}\right)$。$\boldsymbol{U}$ 与 $\boldsymbol{V}$ 具有对称性，$p(\boldsymbol{V}_j\mid\boldsymbol{R},\boldsymbol{U},\boldsymbol{B},\boldsymbol{\Theta}_{\boldsymbol{V}},\alpha)$ 与式(10)具有相同的形式，但 B 的条件后验概率与式(10)略有差异：

$$\begin{aligned}&p\left(\boldsymbol{B}_i\middle|\boldsymbol{R},\boldsymbol{U},\boldsymbol{V},\boldsymbol{\Theta}_{\boldsymbol{B}},\alpha\right)\\&\propto p\left(\boldsymbol{R}\middle|\boldsymbol{B}_i,\boldsymbol{U},\boldsymbol{V},\alpha\right)p\left(\boldsymbol{B}_i\middle|\boldsymbol{\Theta}_{\boldsymbol{B}}\right)\\&\propto\prod_{j=1}^{M}\left[\boldsymbol{\mathcal{N}}\left(\boldsymbol{R}_{ij}\middle|\boldsymbol{B}_i g\left(\boldsymbol{U}_i^{\mathrm{T}}\boldsymbol{V}_j\right),\alpha^{-1}\right)\right]^{I_{ij}}\\&\cdot\boldsymbol{\mathcal{N}}\left(\boldsymbol{B}_i\middle|\boldsymbol{\mu}_{\boldsymbol{B}},\boldsymbol{\Lambda}_{\boldsymbol{B}}\right)\\&\propto\boldsymbol{\mathcal{N}}\left(\boldsymbol{B}_i\middle|\boldsymbol{\mu}_{\boldsymbol{B}i}^{*},\left[\boldsymbol{\Lambda}_{\boldsymbol{B}i}^{*}\right]^{-1}\right)\end{aligned} \tag{11}$$

其中 $\boldsymbol{\Lambda}_{\boldsymbol{B}i}^{*}=\boldsymbol{\Lambda}_{\boldsymbol{B}}+\alpha\sum_{j=1}^{M}\left[g(\boldsymbol{U}_i^{\mathrm{T}}\boldsymbol{V}_j)^2\right]^{I_{ij}}$，$\boldsymbol{\mu}_{\boldsymbol{B}i}^{*}=\left[\boldsymbol{\Lambda}_{\boldsymbol{B}i}^{*}\right]^{-1}$

$\cdot\left(\alpha\sum_{j=1}^{M}\left[g(\boldsymbol{U}_i^{\mathrm{T}}\boldsymbol{V}_j)\boldsymbol{R}_{ij}\right]^{I_{ij}}+\boldsymbol{\Lambda}_{\boldsymbol{B}}\boldsymbol{\mu}_{\boldsymbol{B}}\right)$。

已知样本 $\boldsymbol{B}$，超参数 $\boldsymbol{\Theta}_{\boldsymbol{B}}$ 的条件后验概率可以利用高斯-威沙特分布的性质得到。

$$\begin{aligned}
&\because p\left(\boldsymbol{\mu}_{\boldsymbol{B}},\boldsymbol{\Lambda}_{\boldsymbol{B}}\left|\boldsymbol{\Theta}_0\right.\right)\\
&\quad=\boldsymbol{\mathcal{N}}\left(\boldsymbol{\mu}_{\boldsymbol{B}}\left|\boldsymbol{\mu}_1,(\beta_0\boldsymbol{\Lambda}_{\boldsymbol{B}})^{-1}\right.\right)\mathcal{W}\left(\boldsymbol{\Lambda}_{\boldsymbol{B}}\left|\boldsymbol{W}_1,\nu_1\right.\right)\\
&\therefore p\left(\boldsymbol{\Lambda}_{\boldsymbol{B}}\left|\boldsymbol{B},\boldsymbol{\Theta}_0\right.\right)\ =\mathcal{W}\left(\boldsymbol{\Lambda}_{\boldsymbol{B}}\boldsymbol{W}_{\boldsymbol{B}}^{*},\nu_1+N\right)\\
&p\left(\boldsymbol{\mu}_{\boldsymbol{B}}\left|\boldsymbol{\Lambda}_{\boldsymbol{B}},\boldsymbol{B},\boldsymbol{\Theta}_0\right.\right)=\boldsymbol{\mathcal{N}}\left(\boldsymbol{\mu}_{\boldsymbol{B}}\left|\frac{\beta\boldsymbol{\mu}_1+N\overline{B}}{\beta_0+N},\frac{(\boldsymbol{\Lambda}_{\boldsymbol{B}})^{-1}}{\beta_0+N}\right.\right)\\
&\therefore p\left(\boldsymbol{\Theta}_{\boldsymbol{B}}\left|\boldsymbol{B},\boldsymbol{\Theta}_0\right.\right)=p\left(\boldsymbol{\mu}_{\boldsymbol{B}}\left|\boldsymbol{\Lambda}_{\boldsymbol{B}},\boldsymbol{B},\boldsymbol{\Theta}_0\right.\right)p\left(\boldsymbol{\Lambda}_{\boldsymbol{B}}\left|\boldsymbol{B},\boldsymbol{\Theta}_0\right.\right)\\
&\quad=\boldsymbol{\mathcal{N}}\left(\boldsymbol{\mu}_{\boldsymbol{B}}\left|\frac{\beta_0\boldsymbol{\mu}_1+N\overline{B}}{\beta_0+N},\frac{(\boldsymbol{\Lambda}_{\boldsymbol{B}})^{-1}}{\beta_0+N}\right.\right)\\
&\qquad\cdot\mathcal{W}\left(\boldsymbol{\Lambda}_{\boldsymbol{B}}\left|\boldsymbol{W}_{\boldsymbol{B}}^{*},\nu_1+N\right.\right)
\end{aligned}\tag{12}$$

其中 $\left[\boldsymbol{W}_{\boldsymbol{B}}^{*}\right]^{-1}=\boldsymbol{W}_1^{-1}+N\overline{S}+\frac{\beta_0 N}{\beta_0+N}\left(\boldsymbol{\mu}_1-\overline{B}\right)^2,\overline{B}=\frac{1}{N}=\sum_{i=1}^{N}\boldsymbol{B}_i,\overline{S}=\frac{1}{N}\sum_{i=1}^{N}\boldsymbol{B}_i^2$。

同理，超参数 $\boldsymbol{\Theta}_{\boldsymbol{U}},\boldsymbol{\Theta}_{\boldsymbol{V}}$ 的后验概率与式(12)具有相同的形式。至此，各个参数的条件概率都已经推导出了明确的形式。

3.3 算法复杂度分析

计算 $d\times d$ 逆矩阵的复杂度为 $O(d^3)$，所以 L-BPMF 总的计算复杂度为 $O\left(kd^3\sum_{i=1}^{N}\sum_{j=1}^{M}I_{ij}\right)$，与 BPMF 算法的复杂度相同[11]。RSVD (Regularization SVD)算法的复杂度只有 $O\left(kd\sum_{i=1}^{N}\sum_{j=1}^{M}I_{ij}\right)$，SVD++算法的复杂度为 $O\left(kd\sum_{i=1}^{N}\left(\sum_{j=1}^{M}I_{ij}\right)^2\right)$，但由于贝叶斯模型(BPMF, L-BPMF)提取潜在信息的能力强，往往能用较小的特征维度 d 获得更高的预测准确性。

4 实验设计及结果分析

4.1 数据集合及评价标准

为了测试 L-BPMF 算法的有效性，本文采用推荐系统常用的两种数据集合：Netflix 和 MovieLens。Netflix 数据集合是 Netflix Prize 比赛中使用的标准测试数据集，本文从中随机抽取了包含 8662 名用户对 3000 部视频的约 3×10^5 条评分信息(评分密度为 1.1%)作为测试集合。MovieLens 数据集合由 GroupLens 提供，MovieLens 1M 数据集包含了 6039 名用户对 3883 部电影的 10^6 条评分信息(评分密度为 4.3%)；MovieLens 100K 数据集包含了 943 名用户对 1682 部电影的 10^5 条评分信息(评分密度为 6.3%)。

文献[18]全面总结了推荐系统领域各种不同的评价标准，文中采用检验推荐算法最常用的预测误差 MAE 和 RMSE 作为评价依据，预测误差越小则表示算法性能越好。

$$\left.\begin{aligned}
\mathrm{MAE}&=\left|S_{\mathrm{test}}\right|^{-1}\sum_{(i,j)\in S_{\mathrm{test}}}\left|\boldsymbol{U}_i^{\mathrm{T}}\boldsymbol{V}_j-\boldsymbol{R}_{ij}\right|\\
\mathrm{RMSE}&=\sqrt{\left|S_{\mathrm{test}}\right|^{-1}\sum_{(i,j)\in S_{\mathrm{test}}}\left\|\boldsymbol{U}_i^{\mathrm{T}}\boldsymbol{V}_j-\boldsymbol{R}_{ij}\right\|^2}
\end{aligned}\right\}\tag{13}$$

其中 S_{test} 是测试集合，$|S_{\mathrm{test}}|$ 是 S_{test} 中的元素个数。

4.2 实验设计及结果

本文在上述两种数据集合上，以 MAE 和 RMSE 为评价标准，设计了 3 组实验从不同方面对 L-BPMF 的性能进行测试，每组结果都是 10 次实验结果的平均值。

A 组实验在 Netflix 和 MovieLens 1M 数据集合上，随机选取不同比例的训练集合，对 L-BPMF 和经典 BPMF 进行了对比(特征维度 $d=10$)，使用 RSVD(学习率 $\mathrm{lr}=0.005$，正则化因子 $\lambda=0.02$ [3])的预测误差作为参考。实验结果如图 3，图 4 所示。在 Netflix 数据集合上，L-BPMF 比 BPMF 的预测误差小约 1%，而且在训练比例较低时优势更明显，在训练比例为 20%时能够降低约 1.5%；在 MovieLens 1M 数据集合，L-BPMF 比 RSVD 的预测误差小 1%～2%，L-BPMF 和 BPMF 预测准确性基本一致，在训练比例较低时 L-BPMF 略好一点。分析发现，Netflix 数据集合的评分密度只有 1.1%，而 MovieLens 1 M 的评分密度为 4.3%。一种可能的解释是，评分密度较大时 L-BPMF 与 BPMF 的性能基本一致，评分较稀疏时 L-BPMF 比 BPMF 的预测误差更小。

为了验证上述结论，利用 MovieLens 100K 数据集合产生评分密度<1.5%的训练环境，进行 B 组测试。实际应用中的评分密度都在 1%以下，稀疏条件下的实验更能反映算法提取潜在信息的能力。从图 5 中可以看出，L-BPMF 的 RMSE 预测误差比 BPMF 低约 2%，在最坏的情况下也能够降低约 1%的预测误差。这说明 L-BPMF 提取潜在信息的能力要大于 BPMF，能够有效缓解数据稀疏性问题。

C 组以 MAE 为指标对比了 L-BPMF 和其它主流推荐算法(RSVD, SVD++, KNN, Slope One[19])，实验结果如图 6 所示[20]。其中，RSVD 的参数与 A 组实验相同；SVD++算法中学习率 $\mathrm{lr}=0.005$，正则化因子 λ_1=0.015, $\lambda_2=0.005$ [4,14]; RSVD, SVD++的特征空间维度 d 为 20, L-BPMF 算法的特征维度为 10。都只记录了训练过程中的最优值，并未记录

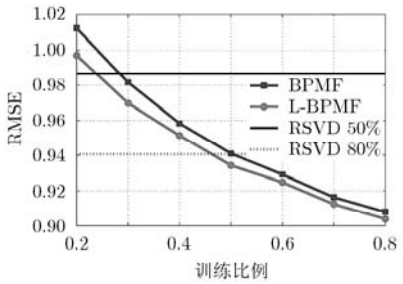

图 3　Netflix 上 L-BPMF 与 BPMF 的对比

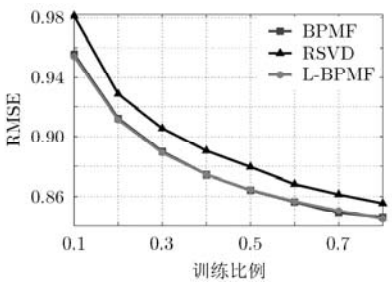

图 4　MovieLens 1M 上 L-BPMF 与 BPMF 的对比

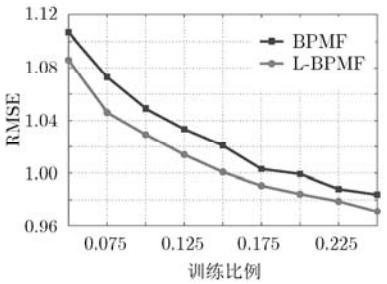

图 5 稀疏条件下 L-BPMF 与 BPMF 的对比

过拟合现象。KNN 算法中以皮尔森相似度公式计算相似性，选择与目标用户最相似的 20 个用户作为邻居进行预测。一般来说，KNN 算法中邻居数目小于 20 时预测准确性会明显下降[14]。

从图 6 中可以看出，L-BPMF 和 SVD++是预测误差最小的两种算法，L-BPMF 的 MAE 误差比 RSVD 低约 0.5%∼1.5%。因为 L-BPMF 特征维度为 10, SVD++的特征维度为 20，而且 L-BPMF 的性能略高于 SVD++，这说明 L-BPMF 能够用较小的特征维度 d 获得更高的预测准确性。

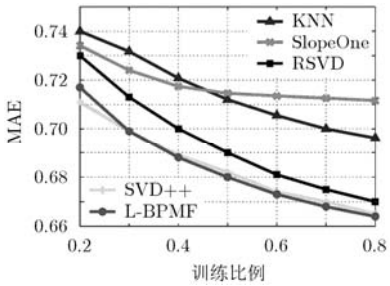

图 6 主流推荐算法的 MAE 预测误差比较

5　结束语

本文针对经典贝叶斯概率矩阵分解模型不能表征潜在因子间非线性关系的问题，提出一种利用 Logistic 函数的 L-BPMF 模型，并使用 MCMC 方法对模型进行训练。在两种真实数据集合上的实验表明，L-BPMF 能够明显提高预测准确性，用较小的特征维度获得比 RSVD 和 SVD++更好的性能。在稀疏条件下的测试结果表明，L-BPMF 比 BPMF 提取信息的能力更强，能够有效缓解数据稀疏性问题。

参 考 文 献

[1] Koren Y, Bell R, and Volinsky C. Matrix factorization techniques for recommender systems[J]. *IEEE Computer*, 2009, 42(1): 30–37.

[2] Billsus D and Pazzani M J. Learning collaborative information filters[C]. Proceedings of International Conference on Machine Learning, San Francisco, 1998: 48–55.

[3] Paterek A. Improving regularized singular value decomposition for collaborative filtering[C]. Preceedings of KDD Cup and Workshop, California, 2007: 39–42.

[4] Koren Y. Factorization meets the neighborhood: a multifaceted collaborative filtering model[C]. Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, 2008: 426–434.

[5] Lee D and Seung H. Algorithm for non-negative matrix factorization[C]. Proceedings of Advances in Neural Information Processing Systems, Denver, 2000: 556–562.

[6] Salakhutdinov R and Mnih A. Probabilistic matrix factorization[J]. *Advances in Neural Information Processing Systems*, 2008, 20(1): 1257–1264.

[7] Salakhutdinov R and Mnih A. Bayesian probabilistic matrix factorization using Markov chain Monte Carlo[C]. Proceedings of the 25th International Conference on Machine Learning, New York, 2008: 880–887.

[8] Gönen M. Predicting drug–target interactions from chemical and genomic kernels using Bayesian matrix factorization[J]. *Bioinformatics,* 2012, 28(18): 2304–2310.

[9] Zhou T, Shan H, Banerjee A, *et al.*. Kernelized Probabilistic Matrix Factorization: exploiting graphs and side information [C]. Proceedings of SIAM International Conference on Data Mining, California, 2012: 403–414.

[10] Liu Q, Wang C, and Xu C. A modified PMF model incorporating implicit item associations[C]. Proceedings of Tools with Artificial Intelligence (ICTAI), Athens, 2012: 1041–1046.

[11] Zhong E, Fan W, and Yang Q. Contextual collaborative filtering via hierarchical matrix factorization[C]. Proceedings of SIAM International Conference on Data Mining, California, 2012: 744–755.

[12] Mackey L, Talwalkar A, and Jordan M I. Divide-and-conquer matrix factorization[C]. Proceedings of the 25th NIPS,

Granada, 2012: 1134–1142.

[13] Lee J, Kim S, Lebanon G, *et al*.. Local low-rank matrix approximation[C]. Proceedings of the 30th International Conference on Machine Learning (ICML-13), Atlanta, 2013: 82–90.

[14] Cacheda F, Carneiro V, Fernandez D, *et al*.. Comparison of collaborative filtering algorithms: limitations of current techniques and proposals for scalable, high-performance recommender systems[J]. *ACM Transactions on Web*, 2011, 5(2): 1–33.

[15] Lü Lin-yuan, Medo M, Yeung C H ,*et al*.. Recommender systems[J]. *Physics Reports*, 2012, 1(3): 159–172.

[16] Jordan M I. Why the logistic function? a tutorial discussion on probabilities and neural networks[J]. *MIT Computational Cognitive Science Report* , 1995, 9503: 1–13.

[17] 仇德辉. 数理情感学[M]. 长沙: 湖南人民出版社, 2001: 55–60.
Qiu De-hui. Mathematical Emotions[M]. Changsha, Hunan People's Publishing House, 2001: 55–60.

[18] 朱郁筱, 吕琳媛. 推荐系统评价指标综述[J]. 电子科技大学学报, 2012, 41(2): 163–175.
Zhu Yu-xiao and Lü Lin-yuan. Evaluation metrics for recommender systems[J]. *Journal of University of Electronic Science and Technology of China*, 2012, 41(2): 163–175.

[19] Lemire D and Maclachlan A. Slope one predictors for online rating-based collaborative filtering[J]. *Society for Industrial Mathematics*, 2005, 5(1): 471–480.

[20] 方耀宁, 郭云飞, 丁雪涛, 等. 一种基于局部结构的改进奇异值分解推荐算法[J]. 电子与信息学报, 2013, 35(6): 1284–1289.
Fang Yao-ning, Guo Yun-fei, Ding Xue-tao, *et al*.. An improved Singular Value Decomposition recommender algorithm based on local structures[J]. *Journal of Electronic & Information Technology*, 2013, 35(6): 1284–1289.

方耀宁： 男，1987 年生，硕士生，研究方向为社会化网络、推荐系统.

郭云飞： 男，1963 年生，教授，博士生导师，研究方向为高性能交换技术、网络安全.

兰巨龙： 男，1962 年生，教授，博士生导师，研究方向为宽带信息网络、下一代互联网.