第49卷第7期 中国科学技术大学学报 Vol. 49, No. 7
2019年7月 JOURNAL OF UNIVERSITY OF SCIENCE AND TECHNOLOGY OF CHINA Jul. 2019

文章编号:0253-2778(2019)07-0572-07

基于辅助信息的混合线性矩阵补全模型

宋 辉,杨 明

(南京师范大学计算机科学与技术学院,江苏南京 210046)

摘要:矩阵补全技术在近年来已经在诸多领域得到了应用,利用已有的辅助信息进行矩阵补全提高补全的精度得到了关注.鉴于此,提出一种将双线性关系与单边线性关系混合的矩阵补全模型,同时关注行信息与列信息之间的相关性和他们各自分别做具有的特点,使得混合线性模型尽可能地逼近原始观测矩阵.;同时证明了使用 ADMM 算法求解的收敛性,并通过拟合数据和真实数据两组实验进行了验证,结果表明,与其他使用辅助信息的补全模型相比,该方法获得补全结果在 RMSE 评价标准下的误差相对降低了 25%以上.

关键词:矩阵补全;辅助特征信息;混合线性;行列相关性;ADMM 算法

中图分类号:TP391 **文献标识码**:A doi:10.3969/j.issn.0253-2778.2019.07.007

引用格式:宋辉,杨明. 基于辅助信息的混合线性矩阵补全模型[J]. 中国科学技术大学学报,2019,49(7):572-578.

SONG Hui, YANG Ming. Mixed linear matrix completion model based on auxiliary information[J]. Journal of University of Science and Technology of China, 2019,49(7):572-578.

Mixed linear matrix completion model based on auxiliary information

SONG Hui, YANG Ming

(*Department of Computer Science and Technology, Nanjing Normal University, Nanjing* 210046,*China*)

Abstract: The matrix completion technology has been applied in many fields in recent years. Using existing auxiliary information to perform matrix completion to improve the accuracy of the completion has attracted attention. A matrix completion model is proposed, which mixes bilinear and unilateral linear relationships, considering the correlation between row information and column information and their respective characteristics, so that the mixed linear model can approximate the original matrix entries. At the same time, the convergence of using the ADMM algorithm to solve the convex optimization problem is proved, and makes two sets of experiments with synthetic datasets and real datasets, which proves that the proposed method is more effective compared with the existing model using auxiliary information, whose error under RMSE evaluation standard was reduced by more than 25% than other methods.

Key words: matrix completion; auxiliary feature information; mixed linear; row and column correlation; ADMM method

收稿日期:2018-06-22;**修回日期**:2018-07-31

作者简介:宋辉,男,1994 年生,硕士生.研究方向:机器学习、推荐系统. E-mail: songhui94@163.com

通讯作者:杨明,男,博士/教授, E-mail:myang@njnu.edu.cn

0　引言

矩阵补全近年来已经成为计算机视觉计算机视觉[1]、推荐系统[2-3]、信号处理[4-5]、社交网络分析[6]等诸多应用领域的一项重要研究内容，经典的低秩矩阵补全模型为经典的基于核范数[7]和矩阵分解[8]的矩阵补全模型，其中标准的矩阵补全问题是描述如下：假设未知矩阵 $M \in R^{m\times n}$ 是低秩的，其观测到的样本元素的集合为 $\{M_{i,j}:(i,j)\in\Omega\}$，则标准的矩阵补全问题描述如下[9]：

$$\lim_{X\in R^{m\times n}} \operatorname{rank}(X) \text{ s.t. } P_\Omega(M)=P_\Omega(X) \quad (1)$$

式中，$\Omega\subseteq[m]\times[n]([m]=\{1,2,\cdots,m\}[n]=\{1,2,\cdots,n\})$ 是采样观察到的元素索引集合，$P_\Omega(\cdot)$ 是正交投影算子，表示当 $(i,j)\in\Omega$ 时，$M_{i,j}$ 为观测数据，即

$$P_\Omega(M_{i,j})=\begin{cases}M_{i,j}, & \text{if } (i,j)\in\Omega\\ 0, & \text{otherwise}\end{cases}.$$

最近一个新的可行的研究方向是基于边缘辅助信息的矩阵补全方法，例如在著名的 Netflix 问题中，除了给定的用户评分信息外，还会有给定的用户信息和关于电影的一些特征信息，若能够利用这些额外的信息，将这些信息加入矩阵补全问题中，那么将可能会得到更好的预测和推荐效果.

一些研究已经开始探索将辅助信息融入矩阵补全的问题中，并用实验证明了辅助信息在一些特定的应用领域对最终的预测结果是确实有效的[10-11]. 例如，为了预测用户—电影评分矩阵中确实的评分，就可以根据已知的用户描述和电影描述信息，将其加入基于内容的推荐系统中，极大地改善了其预测效果. 例如，儿童或年轻人可能对卡通漫画有更高的选择倾向，那么用户的年龄信息可能就会与卡通类型的影片具有相互作用，在此情况下，当用户评分稀疏时，辅助信息就能为补全提供帮助. 有研究证明[12-13]，假如有了非常适合的特征辅助信息，那么在一定条件下可以通过解决嵌入特征的方式来降低样本的复杂度，从而对矩阵补全提供非常巨大的帮助. 由于该模型对特征条件的适合性要求严格，多数场景无法达到标准，从而一个通用的更适合多数场合下的模型显得更加重要.

本文研究了如何通过辅助信息来进行矩阵补全问题，主要贡献有：提出了一种双线性与单边线性关系融合的方式，不仅考虑到用户与电影之间的相互关系，也探讨了在用户本身的特征信息和电影本身所具有的信息会对评分有怎样的影响. 实验结果证明，在考虑了双线性与单边线性关系之后，基于辅助信息的矩阵补全具有更好的补全效果.

1　相关工作

基于压缩感知[14-15]发展的矩阵补全理论已经被广泛应用于许多的机器学习任务中，如视频去噪[16]、图像恢复[17-18]、数据聚类[19-20]、网络服务属性监测[21]等领域. 在此基础上，建立了一些理论，其中一项重要的成果是 Candès 等[22]为矩阵补全提供的重要理论依据，证明了在 $O(n\log n)$ 个观测值的情况下，且样本是均匀随机采样的，此时矩阵能够被准确恢复. 最近一些研究已经开始考虑将辅助信息加入矩阵补全的问题中，研究证明，这些辅助特征信息在某些应用[12,23]和冷启动[13]问题中确实起到了一些作用，然而这些研究大多集中在非凸的矩阵分解模型上，且没有理论保证. 基于凸优化的核范数正则化模型近年来也曾提出[24]，并且为基于辅助信息的矩阵补全提供了理论依据，这些模型都建立在双线性模型 $X^{\mathrm{T}}GY$ 上，且约束关系为 $R_\Omega(X^{\mathrm{T}}GY)=R_\Omega(M)$，其中 X 和 Y 分别表示待补全(评分)矩阵的行记录(如用户)和列记录(如电影)的特征信息. 这种双线性关系模型的参数矩阵 G 或者要求是低秩的[25]或者要求其核范数 $\|G\|_*$ 最小[26]. 在这种模型中恢复一个较小规模的 G 要比直接恢复矩阵 M 要简单得多. 在强假设完美的辅助信息时，即 X 和 Y 都是正交矩阵，且分别在矩阵 M 的列与行的隐空间内，本文提出的方法能够在大幅减少样本复杂度的情况下，准确恢复矩阵 M. 在实际过程中，由于辅助特征信息 X 与 Y 并不是那么完美，最近一系列研究[27-28]提出用残差矩阵 N 解决辅助信息带有噪声的问题. 该方法构建了 $X^{\mathrm{T}}GY+N$ 来逼近 M，并同时要求 G 和 N 是低秩的或有一个较小的核范数. 该方法用 $\|N\|_*$ 来刻画辅助信息对矩阵恢复的有效程度，同时证明了在 X 和 Y 包含了整个 M 的隐空间时恢复矩阵所需的样本复杂度为 $O(N)$，在包含部分隐空间时所需的样本复杂度为 $O\log N$.

本文提出了一种新的基于辅助信息的模型 $X^{\mathrm{T}}GY$ 来逼近 M. 其中 G 可以是稀疏的，但并不需要是低秩的，并且 G 的第 (i,j) 项决定了用户的第 i 项特征与产品第 j 项特征间的相互关系. M 的低秩性通常用来表述用户在面临相类似的物品时近似偏好的选择程度. 假如有近似 $X^{\mathrm{T}}GY=M$，则有 $\operatorname{rank}(M)\leqslant\operatorname{rank}(G)$，因此 G 的低秩性成为 M 低秩性的充分条件. 已有的研究都强调了矩阵 G 的低

秩条件,然而对 M 并没有低秩性的要求. 本文中,当给定的辅助信息 $X \in R^{d_1 \times n}$, $Y \in R^{d_2 \times n}$,其中 $d_1, d_2 \ll N$ 时,对矩阵 G 的低秩要求就会变得过强. 本文用一个低秩的矩阵 E 来直接近似矩阵 M,同时用 G 的稀疏正则化在估计 E.

2 基于辅助信息的混合线性矩阵补全

为了利用 X 和 Y 来补全矩阵 M,我们考虑利用双线性与单边线性混合的方式来建立一个预测模型,以通过观测到的数据预测缺失的信息. 我们可以建立一个线性关系:对于用户 i 对物品 j 的评分,假设为 $M_{i,j} = x^{\mathrm{T}} H y + x^{\mathrm{T}} u + y^{\mathrm{T}} v + g$, x 是用户的特征向量,y 是物品的特征向量,u, v, g 为模型参数. 其中, $x^{\mathrm{T}} H y$ 假设用户的某些特征是与物品的某些特征相关联的. 例如,用户与电影的关系,性别为男性,年龄为青年期的用户可能对电影标签为科幻、动作的作品有较大相关性,而年轻女性对与标签为感情、剧情的电影有更多的偏好. $X^{\mathrm{T}} u + y^{\mathrm{T}} v + g$ 的意义为用户本身的某些特征与物品本身的某些特征对评分的影响. 例如,年龄较小的用户可能会给出更高的分数,推理悬疑类电影的受众面较小,一般评分会较低等. 假设

$$\overline{x} = [x \quad 1], \overline{y} = [y \quad 1],$$

$$G = \begin{pmatrix} H & u \\ v^{\mathrm{T}} & g \end{pmatrix} \in R^{(\mathrm{len}(X)+1)\times(\mathrm{len}(Y)+1)}$$

式中, $\mathrm{len}(*)$ 表示参数的维数,那么相对于 DirtyIMC,可以有以下模型:

$$\min_{M} \sum_{(i,j)\in\Omega} l((\overline{X} G \overline{Y}^{\mathrm{T}})_{I,J}, R_{i,j}) + \lambda_G \| G \|_* \quad (2)$$

为了不损失原观测矩阵的信息,将模型改为

$$\min_{M,E} l(G) + \lambda_E \| E \|_*$$
$$\text{s.t. } \overline{X} T G \overline{Y} = E, R_\Omega(E) = R_\Omega(M) \quad (3)$$

式中,E 是观测矩阵 M 的补全结果, $\overline{X}, \overline{Y}$ 是 X 和 Y 的增广矩阵, $l(G)$ 和 $\| E \|_*$ 分别用来量化估计 G 的稀疏性和 E 的低秩性. 因为辅助信息有可能是含有噪声的,也有可能并非所有的辅助特征信息都对补全是有效的,因此 G 的稀疏性是非常有必要的,本文使用 $l(G) = \| G \|_1$,另外 E 的低秩性也自然会有要求,因为其实矩阵 M 的补全结果. 参数 λ_E 用来平衡两部分的权重.

假设忽略迭代计算带来的损失,并且为了便利,将增广辅助特征信息标记为 X, Y,并假设噪声为高斯噪声,我们放宽约束 $X^{\mathrm{T}} G Y = E$,转为约束 $\| X^{\mathrm{T}} G Y - E \|_F^2$ 最小,则可以将原问题转化为下面的凸优化问题:

$$\min_{M,E} \frac{1}{2} \| X^{\mathrm{T}} G Y - E \|_F^2 + \lambda_G \| G \|_1 + \lambda_E \| E \|_*$$
$$\text{s.t. } R_\Omega(E) = R_\Omega(M) \quad (4)$$

相对于已有的模型,本文提出的方法有以下几点不同:已有的模型都通过寻找一个相关矩阵 H 使得当 $\| H \|_*$ 最小时有 $R_\Omega(X^{\mathrm{T}} G Y) = R_\Omega(M)$,本文将其双线性的关系扩展为含有单边线性的模型,同时考虑了线性关系与二次关系,并且可以选择约束 G 稀疏性的范数. 在已有方法中,都会通过选择约束 G 的秩来保证 E(或 M)的低秩,但当 G 的秩被错误估计时, E 的低秩性就无法得到保证. 在本文中,除了 G 的低秩性, X 或 Y 的低秩也能够使得 E 保持低秩. 另一方面,当 λ_G 足够大的时候,原问题就会因为 G 会迭代到零矩阵而退化成一个标准的不含辅助信息的矩阵补全问题,从另一方面说明了在没有或者较少辅助信息时,我们的模型也是适用的.

为了使变量可分离,并且能够使用 ADMM 算法,假设 $C = E - X^{\mathrm{T}} G Y$,则原问题对应的 Lagrangian 函数为

$$L(C, E, G, M_1, M_2, \beta) = \frac{1}{2} \| C \|_F^2 + \lambda_E \| E \|_1 + \lambda_G \| G \|_* + \langle M_1, R_\Omega(E - M) \rangle + \langle M_2, E - X^{\mathrm{T}} G Y - C \rangle + \frac{\beta}{2} \| R_\Omega(E - M) \|_F^2 + \frac{\beta}{2} \| E - X^{\mathrm{T}} G Y - C \|_F^2.$$

式中, M_1, M_2 为拉格朗日参数, $\beta > 0$ 为参数系数. 当给定第 k 次迭代后的 $G^k, E^k, C^k, M_1^k, M_2^k$ 时,则对于每个变量的迭代问题求解分别如下:

$$C^{k+1} = \arg\min_C L(E^k, G^k, M_2^k, C, \beta^k) = \frac{\beta^k}{\beta^{k+1}} \left(E^k - X^{\mathrm{T}} G^k Y + \frac{M_2^k}{\beta^k} \right) \quad (5)$$

$$G^{k+1} = \arg\min_G L(Ek, G, M_2^k, C^{k+1}, \beta^k) = \mathrm{reshape}\left(\max\left(\left| g^k - \frac{f^k}{\tau_A} \right| - \frac{\lambda_G}{\tau_A \beta^k}, 0 \right) \mathrm{sgn}\left(g^k - \frac{f^k}{\tau_A} \right) \right) \quad (6)$$

式中, $f^k = y^{\mathrm{T}} X^{\mathrm{T}} \left(y^{\mathrm{T}} X^{\mathrm{T}} g^k + c^k + e^k - \frac{m_2^k}{\beta^k} \right)$, $e = \mathrm{vec}(E)$, $g = \mathrm{vec}(G)$, $m = \mathrm{vec}(M)$, $c = \mathrm{vec}(C)$, $\mathrm{resahpe}(s)$ 表示把一个向量 $s \in R^{ab}$ 转换成一个矩阵 $S \in R^{a \times b}$,与操作符 $\mathrm{vec}(S)$ 互为逆运算. $A \otimes B$

表示矩阵 A 和 B 的 Kronecker 内积.

$$E^{k+1}=\arg\min_{E}L(E,G^{k+1},M_1^k,M_2^k,C^{k+1},\beta^k)=\mathrm{SVT}\left(E^k-\frac{p^k+q^k}{2\tau_B},\frac{\lambda_E}{2\tau_B\beta^k}\right) \tag{7}$$

式中，$p^k=R_\Omega\left(E^k-M+\frac{M_1^k}{\beta^k}\right)$，$q^k=E^k-X^{\mathrm{T}}G^{k+1}Y-C^k+\frac{M_2^k}{\beta^k}$，SVT($A$,$B$)操作算子为奇异值阈值操作[7]，表示对矩阵 A 的奇异值对角矩阵进行 max(σ_i-t,0)选择操作，σ_i 表示 A 矩阵奇异值分解后的第 i 个奇异值.

$$M_1^{k+1}=M_1^k+\beta^k(R_\Omega(E^{k+1}-M)) \tag{8}$$

$$M_2^{k+1}=M_2^k+\beta^k(E^{k+1}-X^{\mathrm{T}}G^{k+1}Y-C^{k+1}) \tag{9}$$

算法 2.1　mixed linear matrix completion

输入：X,Y,R_Ω(M)，参数 λ_G,λ_E, τ_A,τ_B,ρ,$\beta_{\max}$

输出：C,G,E

初始化 G^0,E^0,C^0,M_1^0,M_2^0,k=0

循环：用(5)式更新 C^{k+1}

用(6)式更新 G^{k+1}

用(7)式更新 E^{k+1}

用(8)式更新 M_1^{k+1}

用(9)式更新 M_1^{k+1}

$\beta_{k+1}=\min(\beta_{\max},\rho\beta_k)$

K=k+1；直到收敛

返回 C,G,E

在初始化的过程中，M_1^0,M_2^0 使用标准高斯分布生成的随机序列，G^0,E^0 分别使用 SVT 方法和 $R_\Omega(E^0)=R_\Omega(M)$ 得到. 同时我们给出以下定义，以保证算法的性能，使得算法能够快速收敛.

定义 2.1　令 $B(E)=\begin{pmatrix}\Omega(E)\\E\end{pmatrix}$，$\mathrm{A}(G)=\begin{pmatrix}0\\-X^{\mathrm{T}}GY\end{pmatrix}$，$M=\begin{pmatrix}M_1\\M_2\end{pmatrix}$，若给定 β_k 是非递减的并且具有上界，$\tau_A>\|A\|^2$，$\tau_B>\|B\|^2$，那么通过算法 2.1 得到的序列 $\{(G^k,E^k,C^k,M^k)\}$ 可以快速收敛到式(4)的全局最优解.

综上，本文提出的算法可有效地解决问题. 首先在经典的分块坐标下降方法[29]（block-wise coordinate descend，BCD）中，算法的收敛性必须由优化问题的严格凸来保证，并且要求每次迭代中交替优化的子问题具有唯一的最优解. 实践中通常是具一个非常强的约束条件，而本文提出的非严格的凸优化问题，可以通过定义 2.1 的收敛性来保证最后的解是全局最优的解. 其次，在子问题中，$L1$ 范数和核范数是非光滑的函数，通常的优化方法很难适用求解得到最优解，本文提出的线性化求解算法能够为每个子问题提供解决方法，并能够提高迭代效率.

3　实验及结果分析

实验采用拟合的模拟数据和现实存在的真实数据 Movielens[30] 和 NCI-DREAM 数据集进行实验，采用对比的是最近也使用了辅助信息的 3 种算法模型：MAXIDE[26]，IMC[25] 及 DirtyIMC[27]. 本文重点展示算法在实践中的有效性，所有的性能通过计算均方误差 RMSE=$\frac{\|R_\varphi(X^{\mathrm{T}}MY-R)\|_2^2}{\|R_\varphi(R)\|_2^2}$ 来衡量，其中 $R_\varphi(A)$ 是矩阵 A 中缺失的元素. 对拟合数据和真实数据，本文首先随机将观测矩阵 M 中的元素按百分比作丢失处理，通过相同的指定数据集中选取随机 10%的记录进行交叉验证；然后将各模型算法应用到该指定的数据集中，得到各算法的平均 RMSE 值，同时调整超参数 λ 分别为 10^{-3}，10^{-2}，…，10^4. 重复上述过程 4 次后取平均值，获得对算法最有利的参数值. 对于算法 IMC 和 DirtyIMC，本文选取 G 的秩从 rank=1 到 rank=15 变化，对缺失元素占数据总数的百分比，通过比较上述的算法最后的 RMSE 获得.

3.1　人造数据

对人造数据，论文设计了 X 和 Y 分别是满秩的与非满秩的两种实验. 我们随机初始化 X 和 Y，为了使模拟效果更符合实际效果，辅助特征信息尽可能是合成的，X 和 Y 中每个特征都是根据高斯分布、泊松分布、伽马分布中随机选择分布生成的，稀疏矩阵 G 的元素是通过高斯分布 $N(0,100)$ 生成，且缺失元素的位置通过随机挑选来决定. 同时使得观测矩阵 $M=X^{\mathrm{T}}GY+N$. 其中，N 表示噪声，每个 $N_{i,j}$ 的值从标准正态分布分布中产生. 对于元素缺失比例，本文设计其从 10%递增到 80%，每次递增 10%.

为了同其他 3 种方法进行比较时，分别在不同的 3 种设置下进行了实验，分别标记为模拟实验 1，2，3. 在拟合实验 1 中，X 和 Y 的维数分别为 15×50

和 20×140，并且两个矩阵都是满秩的. 与此相对，拟合实验 2，3 采用的 X 与 Y 数据都是非满秩的，并且在拟合实验 2 中，X 与 Y 的维数分别为 16×50 和 21×140；实验 3 中，X 与 Y 的维数分别为 20×50 和 25×140，且在实验 2 和 3 中，X 的前 15 个、Y 的前 20 个特征信息是随机建立的，其余的记录都是根据前面的记录线性组合得到的. 对于 3 组实验，我们记录了各方法得到平均 RMSE，结果如图 1 所示.

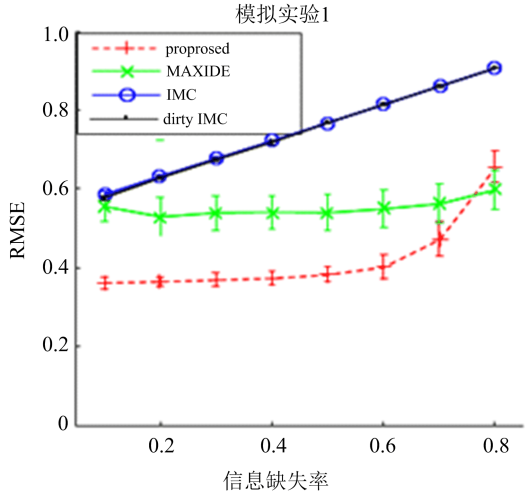

图 1 模拟实验 1

Fig. 1 Simulation experiment 1

从图 1 可以看出，本文提出的方法比其他 3 种方法都有效，并且在数据缺失比例上升时，本文提出方法的 RMSE 增长也比其他方法缓慢. 在研究针对不同的方法，最适合的 G 和 E 的秩的数量时，我们发现，对于本文方法、MAXIDE、IMC、DirtyIMC，G 的秩分别为 15、8、1、1，E 的秩分别为 15、15、1、2 时，对应的算法有最佳的效果. 这些结果表明之前所提到的对于 G 有较强的低秩要求时，算法的性能可能会因此受到影响. 此外，本文研究了恢复后的 G 矩阵与原观测矩阵之间的对比差异，在使用实验 1 的设置、信息缺失率为 10% 的情况下，我们分别绘制了各算法经过补全后的 G 与原观测矩阵 G 作对比，图 2 中深色部分表示数值的绝对值较大. 从图 2 可以看出，本文方法经过补全得到的矩阵最接近原观测矩阵，其他方法只能得到近似的结果.

3.2 真实数据

3.2.1 MovieLens

本文采用的数据为现实世界真实存在的数据，我们使用了 MovieLens 这一数据库. 该数据库可从网络（https://grouplens.org/datasets/movielens/）下载获

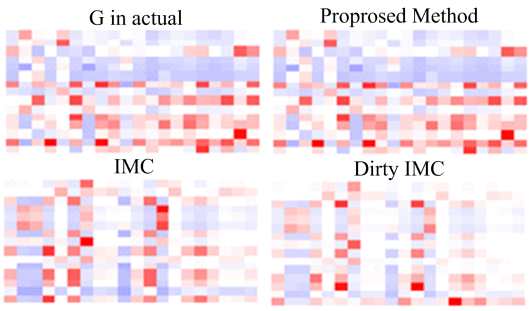

图 2 几种方法得到的 G

Fig. 2 G obtained by several methods

得，本次试验的数据包含 100 000 条用户评分，每条评分记录为整数 1 到 5，一共包括了 943 个用户和 1 682 部电影的信息；并且每一部电影都有 20 条特征信息如类别或是发行时间，每个用户都包含了 24 条特征信息包括年龄性别职业等.

实验分别按信息缺失率为 20%、30%、40%、50%的情况进行，各算法得到的 RMSE 如表 1 所示. 结果显示，本文方法获得的结果比其他几种方法都有效. 图 3 展示了现实中 G 矩阵的一些有规律的数据，如用户在某些数据上表现出相似性，男性在动作、科幻、恐怖、战争类型的电影上会有更高的相关性，但在儿童向的影片偏好较低；退休人员会在假期休闲类型的电影上有着较高的相关度等. 这些相关性在现实中都是具有可解释性的.

表 1 在 MovieLens 上的实验结果

Tab. 1 Result of experimental 1 in MovieLens

RMSE	缺失率			
	20%	30%	40%	50%
Proposed	0.276	0.279	0.284	0.292
MAXIDE	0.424	0.425	0.419	0.421
IMC	0.934	0.943	0.945	0.959
DirtyIMC	0.705	0.738	0.775	0.814

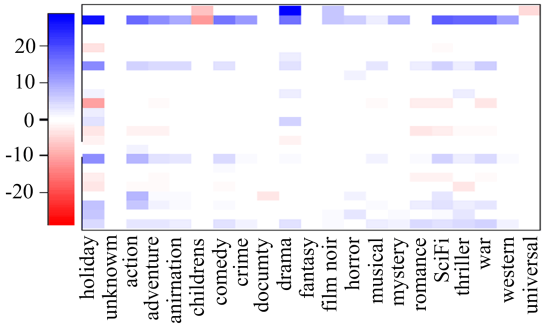

图 3 通过 MovieLens 数据集得到的 G 矩阵

Fig. 3 Matrix G obtained from MovieLens dataset

3.2.2　NCI-DREAM

NCI-DREAM 提供了 46 种乳腺癌细胞系对 26 种药物的反应和所有细胞系的 18 633 种基因的表达数据. 对于每种药物,有 14 个特征描述了它们的化学和物理特性,如分子量,XLogP3 和氢键供体计数,该数据可从网络(http://pubchem.ncbi.nlm.nih.gov/)下载获得.

对于细胞系特征,我们用主成分分析(PCA)选取前 45 个主成分. 实验对 4 种方法进行比较,分别使用了不同的信息缺失率.

表 2 再次证明,在相同的条件下,本文方法具有一定的优越性. 在研究针对不同的方法,最适合的 G 和 E 的秩的数量时发现,对于本文方法、MAXIDE、IMC、DirtyIMC,G 的秩分别为 15、15、1、1,E 的秩分别为 2、15、1、2 时,对应的算法有最佳的效果. 这表明在该数据集上一个低秩的 E 和高秩的 G 会有较好的性能;换句话说,低秩的 G 会使得恢复 E 时的性能降低.

表 2　在 NCI-DREAM 上的实验结果

Tab. 2　Experimental results in NCI-DREAM

RMSE	缺失率			
	20%	30%	40%	50%
Proposed	0.201	0.219	0.225	0.240
MAXIDE	0.268	0.240	0.255	0.288
IMC	0.437	0.489	0.557	0.637
DirtyIMC	0.432	0.475	0.551	0.632

同时实验也绘制了该实验数据的 G 矩阵,如图图 4 所示,列向量表示细胞系的特征,行向量表示药物特征,相关特征可在附录里查看. 根据图 4,我们可以得到:XlogP(F2)、HBD(F3)、HBA(F4)和 rotatable bond number(F5)在药物敏感性测试中起到了重要作用.

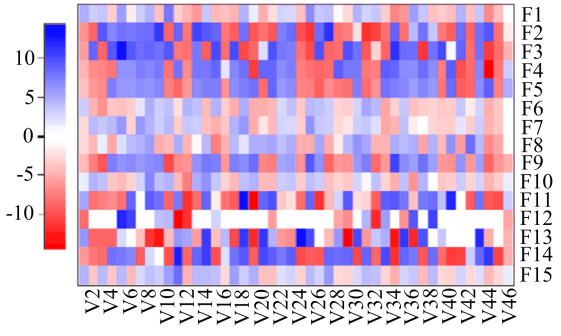

图 4　通过 NCI-DREAM 数据集得到的 G 矩阵

Fig. 4　Matrix G obtained from NCI-DREAM dataset

4　结论

本论文提出了一种基于辅助信息的矩阵补全模型,该模型利用双线性关系与单边线性相结合的方式,利用行与列记录的辅助特征信息进行预测并补全观测矩阵中缺失的记录. 该方法同时使用了线性模型的预测能力和具有相互关系下的双线性模型的预测能力,使得算法具有更好的效果. 同时用理论分析了该算法的收敛性,有别于经典的 BCD 算法解决矩阵补全问题,说明了使用算法 2.1 的解是全局最优解. 最后用实验证明了该方法在拟合数据和真实数据上的表现优于其他 3 种已知的同样利用辅助信息的方法. 在未来的工作中,我们可能会考虑将辅助信息的噪声添加进入模型中以及是否能将研究成果应用到其他相关问题,如多标签学习或半监督聚类学习中.

参考文献(References)

[1] CHEN P, SUTER D. Recovering the missing components in a large noisy low-rank matrix: application to SFM[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2004, 26(8): 1051-1063.

[2] RENNIE J D M, SREBRO N. Fast maximum margin matrix factorization for collaborative prediction[C]// International Conference. Bonn, Germany: DBLP, 2005: 713-719.

[3] SINDHWANI V, BUCAK S S, HU J, et al. One-class matrix completion with low-density factorizations [C]// IEEE International Conference on Data Mining. Sydney, Australia: IEEE Computer Society, 2010: 1055-1060.

[4] NING X, KARYPIS G. Sparse linear methods with side information for Top-N recommendations [C]// Proceedings of the sixth ACM conference on Recommender systems. New York: ACM, 2012: 155-162.

[5] WENG Z, WANG X. Low-rank matrix completion for array signal processing [C]// IEEE International Conference on Acoustics, Speech and Signal Processing. Kyoto, Japan: IEEE, 2012:2697-2700.

[6] CHIANG K Y, HSIEH C J, NATARAJANN, et al. Prediction and clustering in signed networks: a local to global perspective[J]. Journal of Machine Learning Research, 2014, 15(1): 1177-1213.

[7] CAI J F, CANDèS, E J, SHEN Z. A singular value thresholding algorithm for matrix completion[J]. SIAM Journal on Optimization, 2008, 20(4):1956-1982.

[8] RECHT B, FAZEL M, PARRILO P A. Guaranteed minimum-rank solutions of linear matrix equations via

nuclear norm minimization[J]. SIAM Review, 2010, 52(3): 471-501.
[9] CANDÈS E J, TAO T. The power of convex relaxation: Near-optimal matrix completion[J]. IEEE Transactions on Information Theory, 2009, 56(5): 2053-2080.
[10] YI J, ZHANG L, JIN R, et al. Semi-supervised clustering by input pattern assisted pairwise similarity matrix completion[C]// International Conference on Machine Learning. 2013,28(3): 1400-1408.
[11] CHEN Y, JALALI A, SANGHAVIS, et al. Clustering partially observed graphs via convex optimization [J]. Journal of Machine Learning Research, 2014, 15(1): 2213-2238.
[12] MENON A K, CHITRAPURA K P, GARG S, et al. Response prediction using collaborative filtering with hierarchies and side-information[C]// ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Atlanda, USA: ACM, 2011: 141-149.
[13] NATARAJAN N, DHILLON IS. Inductive matrix completion for predicting gene - disease associations [J]. Bioinformatics, 2014, 30(12): i60-i68.
[14] DONOHO D L. Compressed sensing [J]. IEEE Transactions on Information Theory, 2006, 52(4): 1289-1306.
[15] CANDÈS E J, WAKIN M B. An introduction to compressive sampling [J]. IEEE Signal Processing Magazine, 2008, 20(4): 1956-1982.
[16] JI H, LIU C, SHEN Z, et al. Robust video denoising using low rank matrix completion [C]// Computer Vision and Pattern Recognition. San Francisco: IEEE, 2010: 1791-1798.
[17] KOMODAKIS N. Image completion using global optimization[C]// IEEE Computer Society Conference on Computer Vision & Pattern Recognition. New York: IEEE Computer Society, 2006: 442-452.
[18] KORAH T, RASMUSSEN C. Spatiotemporal inpainting for recovering texture maps of partially occluded building facades[J]. IEEE Transactions on Image Processing, 2007, 16(9): 2262-2271.
[19] ZHANG X, SUN F, LIU G, et al. Fast low-rank subspace segmentation [J]. IEEE Transactions on Knowledge & Data Engineering, 2014, 26(5): 1293-1297.
[20] LIU G, LIN Z, YU Y. Robust subspace segmentation by low-rank representation [C]// International Conference on Machine Learning. Haifa, Israel: DBLP, 2010: 663-670.
[21] LEI C, GENG Y, CHEN Z Y, et al. Web services QoS prediction via matrix completion with structural noise[J]. Journal on Communications, 2015, 30(1): 321-326.
[22] CANDÈS E J, RECHT B. Exact matrix completion via convex optimization [J]. Foundations of Computational Mathematics, 2009, 9(6): 717-772.
[23] SHAMIR O, SHALEV-SHWARTZS. Matrix completion with the trace norm: Learning, bounding, and transducing [J]. The JournaL of Machine Learning Research, 2014, 15(1): 3401 - 3423.
[24] LIU G, LI P. Low-rank matrix completion in the presence of high coherence[J]. IEEE Transactions on Signal Processing, 2016, 64(21): 5623-5633.
[25] JAIN P, DHILLONI S. Provable inductive matrix completion [J]. arXiv preprint arXiv: 2013, 1306.0626.
[26] XU M, JIN R, ZHOU ZH. Speedup matrix completion with side information: Application to multi-label learning [J]. Advances in Neural Information Processing Systems, 2013: 2301-2309.
[27] CHIANG K, HSIEH C, DHILLON I. Matrix completion with noisy side information[J]. Advances in Neural Information Processing Systems 2015, 28: 3429 - 3437
[28] YANG E, RAVIKUMAR P. Dirty statistical models [J]. Advances in Neural Information Processing Systems, 2013: 611-619.
[29] TSENG P. Convergence of a block coordinate descent method for nondifferentiable minimization[J]. Journal of Optimization Theory & Applications, 2001, 109(3): 475-494.
[30] HARPER F M, KONSTAN JA. The MovieLens datasets: History and context[M]. ACM, 2015, 5(4): 1-19.

附录 药物对应特征

Label	Feature Name	Label	Feature Name
F1	Molecular Weight(g/mol)	F7	Monoisotopic Mass(g/mol)
F2	XLogP3	F8	Topological Polar Surface Area
F3	Hydrogen Bond Donor Count	F9	Heavy Atom Count
F4	Hydrogen Bond Acceptor Count	F10	Complexity
F5	Rotatable Bond Count	F11	Defined Atom Stereocenter Count
F6	Exact Mass(g/mol)	F12	Undefined Atom Stereocenter Count