软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2015,26(9):2262−2277 [doi: 10.13328/j.cnki.jos.004718]
©中国科学院软件研究所版权所有.
E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

面向时序数据的矩阵分解*

黄晓宇[1,2], 潘 嵘[2], 李 磊[2], 梁 冰[3], 陈 康[3], 蔡文学[1]

[1](华南理工大学 经济与贸易学院,广东 广州 510006)

[2](中山大学 计算机软件研究所,广东 广州 510275)

[3](中国电信股份有限公司 广东研究院,广东 广州 510630)

通讯作者: 黄晓宇, E-mail: echxy@scut.edu.cn

摘 要: 研究一类特殊的矩阵分解问题:对由多个对象在一组连续时间点上产生的数据构成的矩阵 R,寻求把它近似地分解为两个低秩矩阵 U 和 V 的乘积,即 $R\approx U^T\times V$.有为数众多的时间序列分析问题都可归结为所研究问题的求解,如金融数据矩阵的因子分析、缺失交通流数据的估计等.提出了该问题的概率图模型,进而由此导出了其约束优化模型,最终给出了模型的求解算法.在不同的数据集上进行实验验证了该模型的有效性.

关键词: 矩阵分解;时间序列数据;概率图模型;缺失估计;低秩近似

中图法分类号: TP181

中文引用格式: 黄晓宇,潘嵘,李磊,梁冰,陈康,蔡文学.面向时序数据的矩阵分解.软件学报,2015,26(9):2262−2277. http://www.jos.org.cn/1000-9825/4718.htm

英文引用格式: Huang XY, Pan R, Li L, Liang B, Chen K, Cai WX. Matrix factorization for time series data. Ruan Jian Xue Bao/ Journal of Software, 2015,26(9):2262−2277 (in Chinese). http://www.jos.org.cn/1000-9825/4718.htm

Matrix Factorization for Time Series Data

HUANG Xiao-Yu[1,2], PAN Rong[2], LI Lei[2], LIANG Bing[3], CHEN Kang[3], CAI Wen-Xue[1]

[1](School of Economics and Commerce, South China University of Technology, Guangzhou 510006, China)

[2](Software Institute, Sun Yet-San University, Guangzhou 510275, China)

[3](Academy of Guangdong Telecom Company, China Telecom Corporation Limited, Guangzhou 510630, China)

Abstract: The paper studies a matrix factorization problem for time series data, where the target matrix R consists of the equal length time series data generated by a set of objects. The goal is to find two low rank matrices U and V, such that $R\approx U^T\times V$. Many time series analysis problems, such as finance data analysis and missing traffic data imputation, can be reduced to the proposed model. A probabilistic graphical representation for the problem is proposed, and a constrained optimization model from the graphical representation is derived. The solution algorithms for the proposed model is also presented. Empirical studies show that the proposed model is superior to the baselines.

Key words: matrix factorization; time series data; probabilistic graphical model; missing estimation; low rank approximation

矩阵分解(matrix factorization,简称 MF)是机器学习研究中最重要的工具之一,在协同过滤[1−5]、协作排序[6,7]、社会网络分析[8]以及文本分析等领域都有广泛的应用.特别地,在以 Netflix 竞赛为代表的协同过滤问题的研究中,各种基于 MF 模型的实现都取得了很好的结果,显示了这一模型对解决大规模缺失数据的恢复问题的有效性.

在本文的工作中,我们考虑如下问题:考察由 N 个对象(记为 $o_1,o_2,\ldots,o_N$)在 T 个连续时间点(记为 $t_1,t_2,\ldots,t_T$)上生成的时间序列构成的矩阵 R,其中,$R_{i,j}$ 对应对象 o_i 在时间点 t_j 上的取值,我们希望为 R 找到合适的因子矩阵

* 基金项目: 国家高技术研究发展计划(863)(2012AA12A203); 国家自然科学基金(61003140); 国家社会科学基金(13BTJ005)
收稿时间: 2013-01-22; 修改时间: 2014-07-09; 定稿时间: 2014-09-19

U 和 V,使得 $R\approx U^TV$.

本文的工作在缺失数据的恢复研究中具有广泛的应用,如在交通数据采集系统中,为了采集各道路的通行速度数据,通常使用路面上探测车辆的行驶速度来估算道路的平均通行速度,但若有道路在某个时间点上没有探测车经过,则产生了数据缺失的现象,为填充这些数据,我们可以应用本文的工作,根据已采集获得的(不完整的)数据矩阵 R 拟合因子矩阵 U 和 V,进而由乘积 U^TV 得到对缺失速度值的估计;另一个潜在应用的例子是在某些药学实验中,为评估药物对患者的影响,一般需要根据预设的时间间隔多次从志愿者身上抽血检验其体内的血药浓度,为减少志愿者的痛苦和降低检测的费用,对每位志愿者,我们都可以为其随机选择若干时间点,并仅在这些时间点抽血检验,我们把这些采样获得的抽血数据构成矩阵 R,进而可以使用与上述类似的方式,估算志愿者在其他时间点上的缺失的血药浓度.

对基于矩阵分解的缺失数据恢复研究,受 Netflix 竞赛的促进,在过去 10 年间已经有了极大的进展,但这些工作普遍隐含假设了数据的取值独立于外部环境的变化.而本文的目标矩阵 R 中数据则具有显著的时序特征,如在前文的例子中,城市道路在每一个时间点上的通行速度都与前一个时间点的速度相接近,志愿者服药后体内血药浓度在相邻时间点上的测量值也没有显著差异.这些观测结果提示我们在对类似的数据作矩阵分解时,应该把数据在时间轴上的演化效应也考虑在内.为描述数据的时序特征,在本文的工作中,我们首先提出一个适合于描述时序数据的概率图模型,在该模型中,提出使用两种不同的分布来描述时序数据中产生于相邻时间点上的数据间的联系;进而,将相应的导出两种不同的矩阵分解模型 MAFTIS-I(matrix factorization for time series data-I)和 MAFTIS-II;最后,将分别给出对 MAFTIS-I,MAFTIS-II 的求解策略.

本文第 1 节给出用到的主要记号的定义.第 2 节对矩阵分解的相关研究进行简要的总结.第 3 节提出面向时序数据的概率图模型,并将由其导出相应的矩阵分解模型和给出相应的求解策略.第 4 节将把本文模型分别应用于两个不同的数据集——标普 500 股票在一年内每天的开盘价格数据集和国内某城市交通路网上 1 853 个路段在接近 9 000 个连续时间点上的通行速度数据一一进行恢复估计,以检验该模型的表现.最后是对全文工作的总结以及对未来工作的展望.

1 记 号

在本文中,若无特殊说明,所有向量均指列向量.对向量 $V=[v_1,v_2,\ldots,v_N]'\in\mathbb{R}^n$,我们使用$||V||_0$,$||V||_1$ 和$||V||_2$ 顺次表示 V 的 0,1,2 范数,其中,

- $\|V\|_0=\sum_{i=1}^{n}\mathbb{I}(v_i\neq 0)$,即,$V$ 中非 0 分量的个数(这里,$\mathbb{I}(X)$是指示函数,当 X 为真时,取值为 1;否则为 0);
- $\|V\|_1=\sum_{i=1}^{n}|v_i|$;
- $\|V\|_2=\sqrt{\sum_{i=1}^{n}v_i^2}$.

此外,对矩阵 $X\in\mathbb{R}^{n\times m}$,我们使用 X_k 表示 X 的第 k 列,并记 X 的 Frobenius 范数为 $\|X\|_F=\sqrt{\sum_{i=1}^{n}\sum_{j=1}^{m}X_{i,j}^2}$.

2 矩阵分解

矩阵分解模型(也称为因子分解模型)的基本想法:对给定的原始矩阵 R,寻求两个小规模因子矩阵 U 和 V,使得因子矩阵的乘积可以近似地拟合 R(即 $R\approx U^TV$).一般的,这个分解可以由如下的最优化问题获得:

$$\{U,V\}=\arg\min_{U,V} Loss(U^TV,R)+P(U,V) \tag{1}$$

这里,U 和 V 分别是 $d\times M$ 和 $d\times N$ 矩阵.为限制 U 和 V 的规模,通常有 $d<<\min\{M,N\}$.

$Loss(U^TV,R)$是损失函数,用于刻画模型的经验误差,通常可以进一步分解为拟合矩阵中估计值$(U^TV)_{st}$ 与真

实值 R_{st} 之间差异之和,即:

$$Loss(U^TV,R)=\sum_{s=1}^{M}\sum_{t=1}^{N}l(R_{i,j},U_i^TV_j),$$

其中,函数 $l(x,y)$一般取为平方误差损失$(x-y)^2$[2−4,12,13,16−18]或绝对值误差损失$|x-y|$[5].

函数 $P(U,V)$是惩罚因子,用于刻画模型本身的复杂度.在机器学习理论中,Vapnik 提出使用最小描述长度(MDL)作为模型复杂度的通用度量[9],但事实上,MDL 是不可计算的,因此在具体实现中,一般都使用其他度量代替.由于刻画矩阵复杂度的本质属性是它的秩(rank),所以 $P(U,V)$的一个合适选择是 $rank(U^TV)$[10,11],在 Candes 和 Tao 的研究中已经证明:在一定的稀疏性假设下,这一设置能够几乎无损的还原矩阵 R[12,13].但考虑到实际计算的问题,一种更为自然的选择是把 $P(U,V)$取为 U,V 的 Frobenius 范数的平方和,即 $(\alpha\|U\|_{Fro}^2+\beta\|V\|_{Fro}^2)$.由 Fazel[10,11]与 Srebro[5]的工作,当 $\alpha=\beta=\frac{1}{2}$ 时, $(\alpha\|U\|_{Fro}^2+\beta\|V\|_{Fro}^2)$ 与 $rank(U^TV)$的凸包络(convex envelope)相等,所以,这一选择可以看作是对 $rank(V^TU)$的凸近似.

注意到,公式(1)隐含假设了模型对缺失数据的恢复能力仅依赖于由它对已有数据的拟合能力,所以在本质上,这一模式是直推式的(transductive model)[9].众所周知,直推式模型虽然具有很多良好的性质(如算法的泛化性保证等[14]),但却不擅长对数据联系的刻画[5];另一方面,作为产生式模型(generative model)的典型代表,概率图模型[15]虽然能很容易地描述复杂的数据关系,但却难以为其泛化性提供理论保证.对此,Salakhutdinov 及其合作者提出了概率矩阵分解模型(PMF)[3].PMF 通过为 U,V 引入统计先验,实现了直推式模型与产生式模型两者的结合,并在 Netflix 缺失数据恢复问题上取得了成功.然而,PMF 解决的主要是对数据内部特征间联系的刻画,但却未能进一步描述数据之间的依赖关系.为此,Salakhutdinov[16],Adams[17]和 Pirreducible[18]等人都以类似的方式进行了尝试,他们引入了更高层次的先验(即超先验,hyperprior)来描述 U,V 的先验,从而实现了对数据间联系的表示.但这类模型的计算复杂度都相当高,所以只能应用于中小规模的问题求解[17];而对大规模的数据,则普遍需要使用抽样处理的策略[16,18],因而求解效率不高.

3 面向时序数据的矩阵分解

3.1 模 型

对由 N 个对象在 T 个连续时间点上产生的数据构成的矩阵 R,为获得 R 的一个合适的分解 $R\approx U^TV$,我们考虑图 1 所示的概率图模型(MAFTIS).

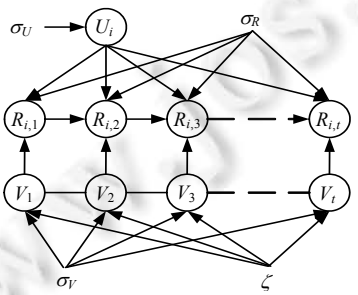

Fig.1 The probabilistic graphical model of MAFTIS

图 1 MAFTIS 的概率图模型

我们假设在图 1 的各组成元素之间存在如下关系:

1) 矩阵 U 中的所有列向量都相互独立地服从于期望为 0、协方差矩阵为 σ_U^2I 的高斯分布,即对于 $1\leqslant i\leqslant N$,有:

$$\Pr(U_i \mid \sigma_U) = (2\pi\sigma_U^2)^{-\frac{d}{2}} \exp\left\{-\frac{\| U_i \|_2^2}{2\sigma_U^2}\right\};$$

2) 矩阵 V 中的所有列向量对于给定的先验 $\sigma_V^2 I$ 与 ζ 构成马尔可夫随机场,即:

$$\Pr(V \mid \sigma_V^2 I, \zeta) = \frac{1}{Z}\prod_{i=1}^{T}\Pr(V_i \mid \sigma_V^2 I, \zeta) \times \prod_{j=2}^{T} Pr(V_j, V_{j-1} \mid \sigma_V^2 I, \zeta);$$

这里, $Z = \int\prod_{i=1}^{T}\Pr(V_i \mid \sigma_V^2 I, \zeta) \times \prod_{j=2}^{T}\Pr(V_j, V_{j-1} \mid \sigma_V^2 I, \zeta)\mathrm{d}V$ 是归一化因子.进一步地,对于 $1 \leqslant j \leqslant T$,我们设 V_j 服从均值为 0、协方差矩阵为 $\sigma_V^2 I$ 的高斯分布,即:

$$\Pr(V_j \mid \sigma_V^2 I, \zeta) = (2\pi\sigma_V^2)^{-\frac{d}{2}} \exp\left\{-\frac{\| V_j \|_2^2}{2\sigma_V^2}\right\}.$$

此外,对于 $\Pr(V_j, V_{j-1} \mid \sigma_V^2 I, \zeta)$,在后文的讨论中,我们分别考虑如下两种情形:

- 情形 I: $\Pr(V_j, V_{j-1} \mid \sigma_V^2 I, \zeta) = \psi_1(V_j - V_{j-1} \mid \zeta)$.这里, $\psi_1(\cdot \mid \zeta)$ 是均值为 0 的拉普拉斯(Laplace)分布,即:

$$\Pr(V_j, V_{j-1} \mid \sigma_V^2 I, \zeta) = \psi_1(V_j - V_{j-1} \mid \zeta) = \frac{1}{2\zeta}\exp\left\{-\frac{\| V_j - V_{j-1} \|_1}{\zeta}\right\};$$

- 情形 II: $\Pr(V_j, V_{j-1} \mid \sigma_V^2 I, \zeta) = \psi_2(V_j - V_{j-1} \mid \zeta)$.其中, $\psi_2(\cdot \mid \zeta)$ 是均值为 0、协方差矩阵为 $\zeta^2 I$ 的高斯分布,即:

$$\Pr(V_j, V_{j-1} \mid \sigma_V^2 I, \zeta) = \psi_1(V_j - V_{j-1} \mid \zeta) = (2\pi\zeta^2)^{-\frac{d}{2}} \exp\left\{-\frac{\| V_j - V_{j-1} \|_2^2}{2\zeta^2}\right\};$$

3) 矩阵 R 中的各成员 $R_{i,j}(1 \leqslant i \leqslant N, 1 \leqslant j \leqslant T)$ 分别独立地服从于期望为 $U_i^T V_j$ 、方差为 σ_R^2 的正态分布,即:

$$\Pr(R_{i,j} \mid U_i^T V_j, \sigma_R^2) = (2\pi\sigma_R^2)^{-\frac{1}{2}} \exp\left\{-\frac{(R_{i,j} - U_i^T V_j)^2}{2\sigma_R^2}\right\}.$$

下面我们根据图 1 和上述的假设 1)~假设 3),对给定的矩阵 R 和参数 $\sigma_U, \sigma_V, \sigma_R, \zeta$,使用极大似然导出本文的矩阵分解模型.需要指出的是:在假设 2)中,我们给出了两种可能的 $\Pr(V_j, V_{j-1} \mid \sigma_V^2 I, \zeta)$ 的分布,由于两者的推导过程类似,我们这里只针对情形 I 进行详细推导,对情形 II 我们将直接给出其主要结论.

在统计意义下,为 R 寻求合适的分解 $R \approx U^T V$ 相当于最大化如下目标:

$$\{U, V\} = \arg\max_{U,V} \Pr(U, V \mid R, \sigma_U, \sigma_V, \sigma_R, \zeta) \tag{2}$$

由

$$\Pr(U, V \mid R, \sigma_U, \sigma_V, \sigma_R, \zeta) = \frac{\Pr(U, V, R \mid \sigma_U, \sigma_V, \sigma_R, \zeta)}{\Pr(R \mid \sigma_U, \sigma_V, \sigma_R, \zeta)} \tag{3}$$

注意到,矩阵 R 和参数 $\sigma_U, \sigma_V, \sigma_R, \zeta$ 均已给定,所以公式(3)中的分母 $\Pr(R \mid \sigma_U, \sigma_V, \sigma_R, \zeta)$ 可视为常数,则公式(2)等价于:

$$\{U, V\} = \arg\max_{U,V} \Pr(U, V, R \mid \sigma_U, \sigma_V, \sigma_R, \zeta) \tag{4}$$

由图 1 和假设 1)~假设 3),我们有:

$$\begin{aligned}
\Pr(R, U, V \mid \sigma_U, \sigma_V, \sigma_R, \zeta) &= \Pr(R \mid U, V, \sigma_U, \sigma_V, \sigma_R, \zeta) \times \Pr(U, V \mid \sigma_U, \sigma_V, \sigma_R, \zeta) \\
&= \Pr(R \mid U, V, \sigma_R) \times \Pr(U \mid \sigma_U) \times \Pr(V \mid \sigma_V, \zeta) \\
&= \prod_{i=1}^{N}\prod_{j=1}^{T}\Pr(R_{i,j} \mid U_i, V_j, \sigma_R) \times \prod_{i=1}^{N}\Pr(U_i \mid \sigma_U) \times \prod_{j=1}^{T}\Pr(V_j \mid \sigma_V) \times \prod_{j=2}^{T}\Pr(V_j, V_{j-1} \mid \zeta). \\
&\propto \exp\left(-\frac{1}{2\sigma_R^2}\sum_{i=1}^{N}\sum_{j=1}^{T}(U_i^T V_j - R_{i,j})^2\right) \times \exp\left(-\frac{1}{2\sigma_U^2}\sum_{i=1}^{T}\| U_i \|^2\right) \times \\
&\quad \exp\left(-\frac{1}{2\sigma_V^2}\sum_{j=1}^{T}\| V_j \|^2\right) \times \exp\left(-\sum_{j=2}^{T}\frac{| V_j - V_{j-1} |}{\zeta}\right)
\end{aligned}$$

对上式两端取对数,有:公式(4)⇔

$$\{U,V\}=\arg\min_{U,V}\frac{1}{\sigma_R^2}\sum_{i=1}^{N}\sum_{j=1}^{T}(R_{i,j}-U_i^TV_j)^2+\frac{1}{\sigma_U^2}\sum_{i=1}^{N}\|U_i\|_2^2+\frac{1}{\sigma_V^2}\sum_{j=1}^{T}\|V_j\|_2^2+\frac{1}{\zeta}\sum_{j=2}^{T}\|V_j-V_{j-1}\|_1 \tag{5}$$

我们令 $\alpha=\frac{\sigma_R^2}{\sigma_U^2},\beta=\frac{\sigma_R^2}{\sigma_V^2},\lambda=\frac{1}{2}\frac{\sigma_R^2}{\zeta}$,则公式(5)可以等价写为模型(I):

$$\left.\begin{aligned}&\{U,V\}=\arg\min_{U,V}\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{T}(R_{i,j}-U_i^TV_j)^2+\frac{\alpha}{2}\sum_{i=1}^{N}\|U_i\|_2^2+\frac{\beta}{2}\sum_{j=1}^{T}\|V_j\|_2^2+\lambda\sum_{j=2}^{T}\|V_j-V_{j-1}\|_1\\&\text{s.t. }\alpha,\beta,\lambda\geqslant 0\end{aligned}\right\} \tag{I}$$

相应地,其优化目标称为目标(I).此外,与上述的过程完全类似,若我们采用假设 2)中的情形 II,我们可以得到以下的模型(II):

$$\left.\begin{aligned}&\{U,V\}=\arg\min_{U,V}\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{T}(R_{i,j}-U_i^TV_j)^2+\frac{\alpha}{2}\sum_{i=1}^{N}\|U_i\|_2^2+\frac{\beta}{2}\sum_{j=1}^{T}\|V_j\|_2^2+\frac{\lambda}{2}\sum_{j=2}^{T}\|V_j-V_{j-1}\|_2^2\\&\text{s.t. }\alpha,\beta,\lambda\geqslant 0\end{aligned}\right\} \tag{II}$$

以下给出模型(I)、模型(II)的性质的若干讨论:

首先,由目标函数的凸性和 KKT 条件[19],我们把模型(I)和模型(II)统一写为模型(III)):

$$\{U,V\}=\arg\min_{U,V}\sum_{i=1}^{N}\sum_{j=1}^{T}(R_{i,j}-U_i^TV_j)^2 \tag{III}$$

s.t.

$$\|U_i\|_2\leqslant A,i=1,2,\ldots,N \tag{C-1}$$

$$\|V_j\|_2\leqslant B,j=1,2,\ldots,T \tag{C-2}$$

$$\|V_{j+1}-V_j\|_q\leqslant C,j=1,2,\ldots,T \tag{C-3}$$

其中,q=1 或 2,A,B,C 是预设常数,分别与α,β和λ这 3 个约束系数相对应:约束系数越大,则相应常数的取值越小;否则反之.

由模型(III)中的约束(C-1)~(C-3)可以看出:在本文的模型中,因子矩阵 U 和 V 的地位是不对称的,我们称 U 为(潜在的)局部对象特征矩阵,$U_1,U_2,\ldots,U_N$ 分别对应数据对象 $o_1,o_2,\ldots,o_N$ 的静态描述;称 V 为(潜在的)全局环境特征矩阵,$V_1,V_2,\ldots,V_T$ 对应了由 $o_1,o_2,\ldots,o_N$ 所共享的外部环境在时间点 $t_1,t_2,\ldots,t_T$ 上的取值.因此,在 MAFTIS 的表示下,矩阵 R 可以看作是由静态的对象特征与动态的环境特征共同作用的结果.如在城市的交通系统中,R 对应了由各道路在不同时间点上的通行速度构成的矩阵,它可以认为是静态的道路特征与时变的环境特征两者的合成.这里的道路特征可能是道路等级、车道数目以及路面平整状况等因素,对所有道路的特征描述构成了因子矩阵 U;另一方面,注意到同一城市内所有道路都共享同样的动态外部环境,如天气状况、能见度、城市人口的生活与工作规律等,对所有时间点上的这些因素的刻画则构成了矩阵 V.

由时序系统在时间上的连续性,对两个相邻的时间点 t_{j-1} 和 $t_j(2\leqslant j\leqslant T)$,当它们之间的间隔充分小时,各数据对象在这两个点上产生的数据应接近相等,即:若$|t_j-t_{j-1}|\to 0$,则在 2 范数意义下,应有:

$$\|R_j-R_{j-1}\|_2\to 0.$$

对于这一行为,约束(C-3)为其提供了保证:

由

$$\|R_j-R_{j-1}\|_2=\|U^TV_j-U^TV_{j-1}\|_2\leqslant\|U\|_2\|V_j-V_{j-1}\|_2,$$

注意到:

$$\|U\|_2\leqslant\|U\|_F\leqslant\sum_{i=1}^{N}\|U_i\|_2\leqslant N\times A.$$

当 q=2 时,令 $C\to 0$,即有:

$$\|R_j-R_{j-1}\|_2\leqslant N\times A\times C\to 0;$$

对 q=1,由于

$$\|V_j-V_{j-1}\|_2\leqslant\|V_j-V_{j-1}\|_1\leqslant C,$$

因此,同样地令 $C\to 0$,我们仍然可以得到$\|R_j-R_{j-1}\|_2\to 0$.

现对模型(I)和模型(II)两者进行对比讨论.虽然通过调节约束(C-3)中 C 的取值(或者相应地,调节模型(I)和模型(II)中的λ值),我们都可以获得 MAFTIS 在时间上的连续性(即,对$|t_j-t_{j-1}|\to 0$,我们可以保证$\|R_j-R_{j-1}\|_2\to 0$),然而模型(I)和模型(II)两者之间仍然有实质性的区别:我们考虑当相邻时间点 t_{j-1} 和 t_j 充分接近时,对全局的环境对象而言,不仅其在 t_{j-1} 和 t_j 这两个时间点上的特征变化幅度趋向于 0,而且环境对象自身发生改变的属性的数目也将趋向于 0.对于前者,我们可以使用 1 范数$\|V_j-V_{j-1}\|_1$ 或 2 范数$\|V_j-V_{j-1}\|_2$ 度量;而对于后者,其精确的取值为$\|V_j-V_{j-1}\|_0$,由于 0 范数的离散性,使其难以适用于大规模的计算.而另一方面,根据 Candes 和 Tao 的结果[20],在一定条件下,$\|X\|_0$ 与$\|X\|_1$ 等价.除此之外,由$\|X\|_2\leqslant\|X\|_1$,所以对于模型(III),在保持 A,B 和 C 的取值不变的情况下,由 q=1,我们可以获得比 q=2 更强的表达能力.

3.2 模型求解

本节讨论对 MAFTIS 的求解,为方便讨论,我们把模型(I)和模型(II)的优化目标合并为如下表达式:

$$S=\arg\min_{U,V}\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{T}(R_{i,j}-U_i^TV_j)^2+\frac{\alpha}{2}\sum_{i=1}^{N}\|U_i\|_2^2+\frac{\beta}{2}\sum_{j=1}^{T}\|V_j\|_2^2+\frac{\lambda}{q}\sum_{j=2}^{T}\|V_j-V_{j-1}\|_q^q \tag{6}$$

这里,q=1 或 q=2.

由于 S 对 $U_i(i=1,2,\ldots,N)$和 $V_j(j=1,2,\ldots,T)$分别为凸,因此我们可以使用交替梯度下降[21]对上式求解.

记 S 对 U_i 的偏导为 $\frac{\partial S}{\partial U_i}$,对 V_j 的偏导为 $\frac{\partial S}{\partial V_j}$,在表 1 中,我们给出了基于公式(6)的矩阵拟合算法.

Table 1 The fitting algorithm

表 1 拟合算法

1	//输入:矩阵 R,潜因子矩阵的维度 d,阈值ε,正则化系数α,β,λ和步长参数η_1,η_2;
2	//输出:潜因子矩阵 U,V.
3	//初始化 U,V.
4	以 $U_1,U_2,\ldots,U_N,V_1,V_2,\ldots,V_T\sim N(0,I)$生成随机矩阵 U 和 V;
5	//交替梯度下降
6	令 $S_1=\arg\min_{U,V}\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{T}(R_{i,j}-U_i^TV_j)^2+\frac{\alpha}{2}\sum_{i=1}^{N}\|U_i\|_2^2+\frac{\beta}{2}\sum_{j=1}^{T}\|V_j\|_2^2+\frac{\lambda}{q}\sum_{j=2}^{T}\|V_j-V_{j-1}\|_q^q$;
7	S_2=inf;
8	While $\|S_2-S_1\|>\varepsilon$
9	$S_2=S_1$;
10	对 i=1,2,…,N, $U_i=U_i-\eta_1\frac{\partial S}{\partial U_i}$;
11	对 j=1,2,…,T, $V_j=V_j-\eta_2\frac{\partial S}{\partial V_j}$;
12	重新计算 S_1;
13	End while
14	输出 U,V.

在表 1 的计算流程中,其核心步骤是对梯度 $\frac{\partial S}{\partial U_i}$ 和 $\frac{\partial S}{\partial V_j}$ 的计算,下面我们分别对此展开讨论.

容易看出,无论 q=1 或 q=2,$\frac{\partial S}{\partial U_i}$ 可使用下式计算:

$$\frac{\partial S}{\partial U_i}=-\sum_{j=1}^{T}(R_{i,j}-U_i^TV_j)V_j+\alpha U_i.$$

下面计算 $\frac{\partial S}{\partial V_j}$.

我们先考察 q=1 的情形,此时,$\|V_j-V_{j-1}\|_q^q=\|V_j-V_{j-1}\|_1$.注意到函数$\|X\|_1$在 0 点不可导,因而需要考虑使用次梯度(subgradient)或其他近似方法.对于前者,注意到在给定因子矩阵 U 的前提下,公式(6)具有与 Fused Lasso 模型[22]相接近的形式,所以一个自然的考虑方向是引入与 Fused Lasso 相类似的解决策略,如基于路径的交替优化模型[23].然而,由于公式(6)对系数向量 V_j 使用了 2 范数惩罚($\|V_j\|_2^2$),这使得文献[23]中的一个依赖于 1 范数惩罚的关键结论——软阈值调节(soft thresholding),不能适用于本文的问题;另一方面,该文提出的交替下降-系数融合(descent-fusion)的策略仍然需要较高的计算代价,因而也不适用于大规模计算的场景.

在本文的工作中,我们考虑如下的近似计算策略:对 $X\in\mathbb{R}^1$,我们定义$\|X\|_1$的基于 2 范数的τ近似为

$$\|X\|_{2\sim\tau}=\sqrt{X^2+\tau}\approx\|X\|_1,$$

其中,τ>0,是一个充分小的预定义常数,这里我们取为 X 的精度的 $\frac{1}{100}$.

进一步地,对向量 $X=[x_1,x_2,\ldots,x_l]'$,我们定义:

$$\|X\|_{2\sim\tau}=[\sqrt{x_1^2+\tau},\sqrt{x_1^2+\tau},\ldots,\sqrt{x_l^2+\tau}]'\approx\|X\|_1.$$

我们在图 2 中展示了当 $X\in[-1,1]$时,在步长为 0.05,τ=0.0001 的设置下,$\|X\|_{2\sim\tau}$对$\|X\|_1$的近似效果,其中,由符号 o 构成的虚线对应函数 $Y=|X|$,由'+'构成的虚线对应了 $Y=\|X\|_{2\sim\tau}$.此外,作为对比,我们也使用连续实线表示了函数 $Y=X^2$ 的曲线.

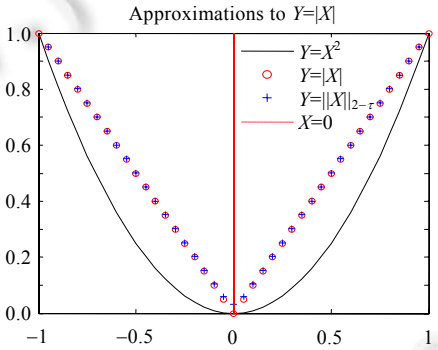

Fig.2 The approximation effects to $\|X\|_{2\sim\tau}$ by $\|X\|_2$ and $\|X\|_1$ respectively

图 2 $\|X\|_{2\sim\tau}$与$\|X\|_2$对$\|X\|_1$的近似效果

从图 2 可以看出,函数 $Y=\|X\|_{2\sim\tau}$与函数 $Y=|X|$两者的曲线几乎完全重合.因而可以预期:当公式(6)中 q=1 时,范数$\|X\|_{2\sim\tau}$与$\|X\|_1$也具有相近的约束能力.

以下我们把公式(6)中的各$\|\cdot\|_1$项都使用为相应的$\|\cdot\|_{2\sim\tau}$近似,进而对 $1\leqslant j\leqslant T$ 计算 $\frac{\partial S}{\partial V_j}$.

为简化表述,我们记 $G_j=-\sum_{i=1}^{N}(R_{i,j}-U_i^TV_j)U_i+\beta V_j$,则:

- 对于 j=1,有:

$$\frac{\partial S}{\partial V_j}=G_j+\lambda\left[-\frac{V_{1,j+1}-V_{1,j}}{\sqrt{(V_{1,j+1}-V_{1,j})^2+\tau}},\ldots,-\frac{V_{d,j+1}-V_{d,j}}{\sqrt{(V_{d,j+1}-V_{d,j})^2+\tau}}\right]';$$

- 对于 1<j<T,有:

$$\frac{\partial S}{\partial V_j}=G_j+\lambda\left[\frac{V_{1,j}-V_{1,j-1}}{\sqrt{(V_{1,j}-V_{1,j-1})^2+\tau}}-\frac{V_{1,j+1}-V_{1,j}}{\sqrt{(V_{1,j+1}-V_{1,j})^2+\tau}},\ldots,\frac{V_{d,j}-V_{d,j-1}}{\sqrt{(V_{d,j}-V_{d,j-1})^2+\tau}}-\frac{V_{d,j+1}-V_{d,j}}{\sqrt{(V_{d,j+1}-V_{d,j})^2+\tau}}\right]';$$

- 对于 $j=T$,有:

$$\frac{\partial S}{\partial V_j}=G_j+\lambda\left[\frac{V_{1,j}-V_{1,j-1}}{\sqrt{(V_{1,j}-V_{1,j-1})^2+\tau}},\ldots,\frac{V_{d,j}-V_{d,j-1}}{\sqrt{(V_{d,j}-V_{d,j-1})^2+\tau}}\right]'.$$

以下计算 $q=2$ 的情形.我们对 G_j 的定义同上,根据公式(6),我们直接有如下结果:

- 对于 $j=1$,有:

$$\frac{\partial S}{\partial V_j}=G_j-\lambda(V_{j+1}-V_j);$$

- 对于 $1<j<T$,有:

$$\frac{\partial S}{\partial V_j}=G_j+\lambda(2V_j-V_{j-1}-V_{j+1});$$

- 对于 $j=T$,有:

$$\frac{\partial S}{\partial V_j}=G_j+\lambda(V_j-V_{j-1}).$$

另外,对于表 1 算法的具体实现,我们还有如下内容的需要注意:

首先,为简化参数的调节,在算法中,我们令 $\eta_1=\eta_2$,并取 $\alpha=\beta$,α 与 λ 的具体取值使用网格搜索策略确定;

其次,表 1 算法的另一个关键问题是对因子矩阵的特征维度 d 的确定,对此,我们首先引述如下结论[24]:

对矩阵 $R\in\mathbb{R}^{N\times T}$ 的一个奇异值分解 $R=U^T\Sigma V$,其中,$U=[u_1,u_2,\ldots,u_N]\in\mathbb{R}^{N\times N}$ 且 $UU^T=I_N$,$V=[v_1,v_2,\ldots,v_T]\in\mathbb{R}^{T\times T}$ 且 $VV^T=I_T$,$\Sigma=diag(\sigma_1,\sigma_2,\ldots,\sigma_{\min\{N,T\}})\in\mathbb{R}^{N\times T}$ 且 $\sigma_1\geqslant\sigma_2\geqslant\ldots\geqslant\sigma_{\min\{N,T\}}$,则对 R 的任意秩数为 k 的近似 R_k,有:

$$\|R-R_k\|_2\geqslant\sum_{i=1}^{k}\sigma_i u_i v_i^T=\sigma_{k+1}.$$

上述结果显示:当 R 中不包含缺失元素时,我们可以根据对 R 的奇异值分解结果和拟合精度的需要选择合适的 d 值.然而当 R 是不完整的矩阵时,由于诸 σ 的取值无法提前获得,因而也难以根据精度的需要确定合适的 d 值.对此,一般地,我们只能以经验的方式选取,即:在实验中调节不同的 d 值,并根据实验结果确定其最终的取值.

4 实 验

4.1 实验设置

- 实验数据

我们把本文算法分别应用于两个不同领域的缺失数据恢复问题中以检验其表现,其中,第 1 个数据集是标普 500 数据集(以下称为 stock 数据集),它使用 524(行)×245(列)的矩阵 R 来记录 2009 年 8 月 21 日~2010 年 8 月 20 日间用于计算标准普尔 500 指数的 524 个组成股票在连续 245 个交易日的开盘价格数据(在此期间,共有 524 个股票被选入了标普 500 指数的计算);第 2 个数据集是城市路网的历史交通速度数据集,它使用 1853(行)×8729(列)的矩阵 R 记录了某城市在连续 30 天内以 GPS 浮动车技术采集获得的市内 1 853 个路段上的平均通行速度数据.这里,R 中的每行对应一个路段,每列对应一个采集时间点,任意两个时间点之间的时间间隔固定为 5 分钟.若路段 i 在时间点 j 上有浮动车经过,则把 $R_{i,j}$ 记为在此时间间隔内路段 i 上所有浮动车速度的平均值;否则,记 $R_{i,j}$ 为空.由于城市内的浮动车数目远小于路段的数目,因而 R 中有大量的数据缺失.事实上, R 中的元素总数目约为 1 600 万,而其中非空元素却不到 800 万个,仅占总数的 50%.

为分析这些数据在时间轴上的变化特点,首先给出数据增长率(rate of increment,简称 roi)的计算定义.

给定 $x,y\in\mathbb{R}$,定义:

$$roi(x,y)=\frac{y-x+c}{x+c} \tag{7}$$

其中,$c\in\mathbb{R}^+$,是一个充分小的常量,用以避免出现分母为 0 的情形.

根据上述定义,我们称 $roi(x,y)$为 y 相对于 x 的增长率.

令 c=0.001,根据公式(7)分别计算了 stock 和 traffic 数据集中各数据对象在相邻时间片上后者相对于前者的数据增长率(即:对每一个矩阵 R,我们都计算了所有的 $roi(R_{i,j-1},R_{i,j})$,这里,$1\leqslant i\leqslant N$ 且 $2\leqslant j\leqslant T$).我们拟合了这些数据增长率的累积密度分布(cumulative density distribution,简称 CDF),其结果如图 3 所示.

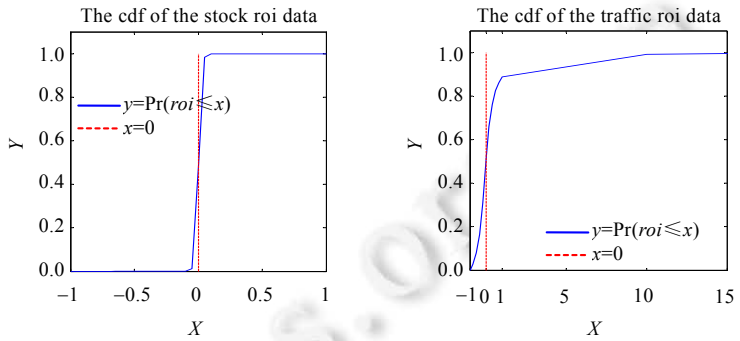

Fig.3 The cdfs of the roi values of the data

图 3 Stock 与 traffic 数据的 roi 值累积密度分布

图 3 左边的子图对应了 stock 数据集,右边的子图对应了 traffic 数据集.为方便对比,我们在两个子图中都同时以虚直线标注了 x=0 的位置.由图 3 可以看出,stock 中各股票每天的开盘价格相对于前一天的增长率都集中在一个非常窄的区间[−10%,10%]之内.这一现象表明:所有股票每天的开盘价格都以前一天的价格为中心,并相对于此价格仅作轻微的波动,从而所有股票在整个时间段上的开盘价格呈现的是渐变的形式.而另一方面,有别于 stock 中的数据,traffic 中各路段通行速度的变化更多的以“突变”的方式出现:在两个相邻时间点之间,增长率在[−10%,10%]间的数据仅占了全体数据的 30%左右,而其他数据的变化幅度则非常大,后一时间点上的路段通行速度既有可能突然降低至 0,也有可能上升到前一时间点的十数倍.由此可以认为:在 traffic 数据集中,各路段在不同时间点上的通行速度是以接近相互独立的方式发生变化的.对于这一现象,我们猜测主要由城市交通自身的特点所决定,由于城市内道路的交通灯较多并且状态转换频密,因此即使对同一路段,它在相邻时间点上的通行速度可能会存在显著的差异.

- 测试协议

对每个数据集,我们在实验中都采用了 Given X 协议[25]对算法进行评估,这里,X 是算法中所使用的训练数据占总体观测数据的比例,余下的数据则作为测试数据用于算法评估.

算法的准确度采用根方平均误差(root mean square error,简称 RMSE)度量,即:

$$RMSE=\sqrt{\frac{1}{|\mathcal{T}|}\sum_{R_{i,j}\in\mathcal{T}}(R_{i,j}-\widehat{R_{i,j}})^2}\ .$$

这里,$\mathcal{T}$是测试数据的集合,$\widehat{R_{i,j}}$ 是对$\mathcal{T}$中数据 $R_{i,j}$ 的估计.

- 基准算法

我们选择概率主成分分析(PPCA)[26]和概率矩阵分解(PMF)[3]作为与本文算法对比的基准算法,其中,前者在交通流缺失数据的恢复研究中获得了非常好的结果[27],后者是在 Netflix 数据集上表现最好的矩阵恢复算法之一.本文实验中的 PPCA 使用了文献[28]的实现,PMF 则采用了文献[29]的实现.

4.2 实验结果

本节用于汇报 MAFTIS 在测试数据集上的结果,包括参数调节对 MAFTIS 的性能的影响、因子矩阵 U,V 的内部结构分析、MAFTIS 在矩阵缺失数据恢复上的表现和 MAFTIS 的收敛速度分析.由于 stock 和 traffic 中的数据在时间轴上的演化模式迥异,所以我们同时汇报了 MAFTIS 在这两个数据集上的运行结果.

- 参数调节

我们先检验对因子α(注意到:在我们的实验设置中,有$\alpha=\beta$)和λ的调节对 MAFTIS 的性能影响.我们分别以一致采样的方式从 stock 和 traffic 两个数据集上抽取了 50%的数据作为测试集,余下的 50%作为训练集,并在 MAFTIS 中分别以固定α、改变λ和固定λ、改变α这两种方式进行数据恢复.在实验中,我们分别把α和λ两者之一固定为 1,另一个则逐步取值为 $2^{-10},2^{-9},\ldots,2^{0},2^{1}$,实验结果如图 4 所示.

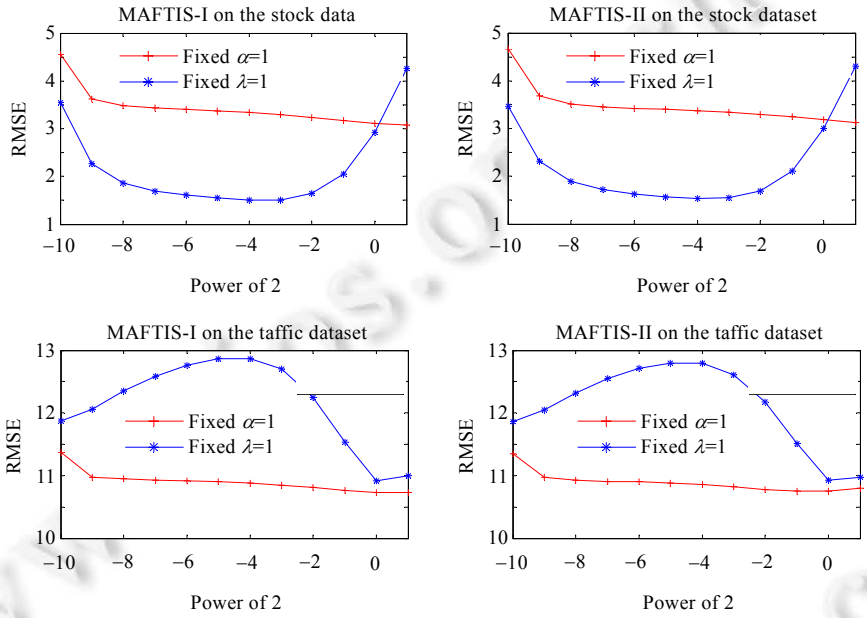

Fig.4 The completion results by MAFTIS with different parameter settings

图 4 MAFTIS 在不同参数设置下的数据恢复结果

图 4 上面一行是 MAFTIS 在 stock 数据集上的数据恢复结果,下一行是在 traffic 数据集上的结果;图中的左列对应模型(I),右列对应模型(II).各图的横坐标对应 2 的幂次,纵坐标对应在(α,λ)参数设置下的 RMSE 值,由'+'构成的折线对应固定α=1,改变λ得到的结果;由'*'构成的折线则对应了固定λ=1,改变α得到的结果.

从图 4 可以看到,模型(I)、模型(II)在相同数据集上的走势几乎完全一致.特别地,在固定α的情况下(对应图中的'+'折线),我们可以看到,公式(6)中的约束项 $\frac{\lambda}{q}\sum_{j=2}^{T}\|V_j-V_{j-1}\|_q^q$ 对预测结果具有显著的影响:在图 4 的 4 个子图中,RMSE 都随着λ的增长稳步降低.值得强调的是:在所有实验的最后(λ=2),我们所获得的预测误差与λ取初值($\lambda=2^{-10}\approx0.001$,此时约束项的影响接近于 0)时相比都有了大幅度的降低.

我们对图 4 结果的另一个有趣发现是:在两个数据集上,预测误差随着α的变化显示出了截然相反的行为:在 stock 数据集上,α越大,则 RMSE 越小;在 traffic 数据集上,RMSE 却随着α的增大而增大.事实上,这一现象可以通过对比假设 1)~假设 3)与公式(6)得到很好的解释:我们固定 σ_R^2 为常数,则有 $\alpha\propto\frac{1}{\sigma_U^2},\beta\propto\frac{1}{\sigma_V^2}$.注意到:stock 中的数据是以"渐变"的行为发生变化的,因而其总体方差 σ_V^2 较小,相应的β应取较大的值;而 traffic 中的数据则以"突变"的行为为主,变化较为激烈,所以其方差 σ_V^2 较大,相应的β的取值应偏小.由于在我们的实现中,为方便参数调节,我们强制规定了$\alpha=\beta$,所以图 4 的现象正是对以上结果的合理反映.

另外,还需指出的是:图 4 显示,在所有情况下,预测误差都稳定的随λ的增加而缩小,因而我们可以通过指数序列 $2^{-10},2^{-9},\ldots,2^{0},2^{1},\ldots,2^{n},\ldots$快速确定$\lambda$的取值.对于$\alpha$(=$\beta$),在 stock 数据集上,其预测误差的"谷底"所对应的α的

取值区间约为[2^{-7},2^{-2}];而在 traffic 数据集上,其预测误差自$\alpha \geqslant 2^{-4}$开始表现出快速下降的趋势,因而α参数的经验最优值也可以通过为其构造 2 的指数序列调节获得.

- 数据恢复结果

下面比较在不同的数据缺失程度下,应用 MAFTIS-I(q=1),MAFTIS-II(q=2)及基准算法对 stock 和 traffic 作数据恢复的效果.对每一个算法,它在每一个数据集上的所有实验都使用相同的参数设置.特别地,对于本文所提出的 MAFTIS-I 和 MAFTIS-II,两者的参数设置完全一致:在 stock 数据集上,$\alpha=\beta=2^{-4}$,λ=2.5;在 traffic 数据集上,$\alpha=\beta$=1,λ=2.5.另外,对于参与实验的其他基准算法,其参数也以上一节参数调节的方式,通过网格搜索确定.

在 Given X 协议中,我们把 X 从 10%开始,按 10%的步长顺次取到 90%,在 stock 数据集上的实验结果汇总在表 2 中,在 traffic 数据集上的结果汇总在表 3 中(结果越小越好).

Table 2 Results on the stock dataset (smaller is better)

表 2 在 stock 数据集上的实验结果(结果越小越好)

		10%	20%	30%	40%	50%	60%	70%	80%	90%
d=10	PPCA	18.52	20.84	24.18	22.82	19.57	18.98	13.93	11.09	6.16
	PMF	3.33	3.29	3.30	3.30	3.28	1.70	1.70	**1.65**	1.78
	MAFTIS-I	**2.84**	**2.08**	2.07	1.95	1.83	1.83	1.81	1.82	1.87
	MAFTIS-II	3.09	**2.08**	**1.99**	**1.93**	**1.75**	**1.69**	**1.66**	1.73	**1.75**
d=30	PPCA	24.22	21.84	24.51	23.22	22.61	21.56	18.23	13.47	10.93
	PMF	3.33	3.29	3.30	3.30	3.29	1.70	1.71	1.65	1.79
	MAFTIS-I	**2.57**	**1.93**	1.96	1.76	1.69	1.63	1.65	1.64	1.70
	MAFTIS-II	3.09	2.02	**1.93**	**1.66**	**1.58**	**1.53**	**1.51**	**1.53**	**1.52**

Table 3 Results on the traffic dataset (smaller is better)

表 3 在 traffic 数据集上的实验结果(结果越小越好)

		10%	20%	30%	40%	50%	60%	70%	80%	90%
d=10	PPCA	17.88	17.36	12.00	12.26	11.47	11.28	11.19	11.17	11.12
	PMF	14.44	12.75	12.49	12.39	12.36	12.35	12.31	12.30	12.26
	MAFTIS-I	12.14	**11.21**	**10.80**	**10.72**	**10.66**	**10.65**	**10.63**	**10.65**	**10.57**
	MAFTIS-II	**12.07**	11.32	11.02	10.86	10.77	10.75	10.73	10.71	10.67
d=30	PPCA	17.89	17.28	13.00	12.24	11.46	11.28	11.19	11.10	11.08
	PMF	14.39	12.77	12.43	12.39	12.35	12.35	12.32	12.31	12.25
	MAFTIS-I	**11.90**	**11.26**	**10.79**	**10.70**	**10.66**	**10.65**	**10.64**	**10.61**	**10.57**
	MAFTIS-II	11.97	11.38	11.07	10.88	10.79	10.76	10.72	10.70	10.67

从表 2、表 3 可以看出:除在 d=10 且 Given 80%这一设置下由 PMF 在 stock 数据集上获得最小的预测误差外,MAFTIS 在其他所有实验中对基准算法都有显著的优势.特别地,在 stock 数据集上,MAFTIS-II 在几乎所有情况下都取得了最好的表现;而在 traffic 数据集上,MAFTIS-I 的优势则更为明显.这启发我们,MAFTIS-I,MAFTIS-II 各有其适用范围:当数据间的变化以渐变方式发生时,MAFTIS-II 可能较为合适;若数据间的变化以突变为主时,则 MAFTIS-I 可能会有更好的结果.

我们还注意到,在 stock 和 traffic 数据集上,调节维度参数 d 对结果的影响也不尽相同:在 stock 数据集上,当 d 值由 10 上升至 30 时,除 PMF 外,其他 3 个算法实现的预测误差都有明显的降低;然而在 traffic 数据集上,这一改变对结果却近乎没有发生影响.对此,我们猜测可能与数据集自身的性质有关:城市交通的实时运行虽然有很大的不确定性,然而总体上却有其周期性,这一宏观规律决定了 traffic 数据集的基空间组成仅需包含少量(≤10)的基向量;而对 stock 数据,由于证券市场变化的周期性并不明显,因而其基空间也需要更多的组成向量,所以当我们提高 d 的取值时,预测结果的质量也相应得到提升.

- 因子矩阵分析

正如我们在前文的讨论中所指出的:MAFTIS 中的矩阵 U 刻画了(独立同分布的)局部数据对象的静态特征,V 则刻画了作用于全局的环境对象在时间上的动态特征.作为对这一结果的直接反映,U,V 内部各相邻列向量之间的差值应有不同的表现:对于 U,由于各列对应的数据对象相互独立,所以相邻列向量之差的分布也没有明显的规律性;而 V 对应了各数据对象在时间轴上的动态行为,因而其内部数据应与 R 具有相同的变化趋势.

为度量 V 中列向量间的变化幅度,我们把向量 Y 相对于 X 的变化率(rate of change)记为 $roc(X,Y)$,其计算方式定义为

$$roc(X,Y)=\frac{\|Y-X\|_2}{\|X\|_2} \tag{8}$$

数据恢复实验中,以 Given 50%设置下获得的因子矩阵为例,分析 U,V 各自的性质.

对于在每一个数据集上的拟合结果,我们分别计算了其 U,V 矩阵内部相邻列间后者相对于前者的变化率(即:$roc(U_i,U_{i+1})$和 $roc(V_j,V_{j+1})$).其中,对 stock 的计算结果如图 5 所示(其中,第 1 行对应 U,第 2 行对应 V;左图由 MAFTIS-1 获得,右图由 MAFTIS-II 获得),对 traffic 数据集的计算结果则如图 6 所示(其中:第 1 行对应 U,第 2 行对应 V;左图由 MAFTIS-1 获得,右图由 MAFTIS-II 获得).

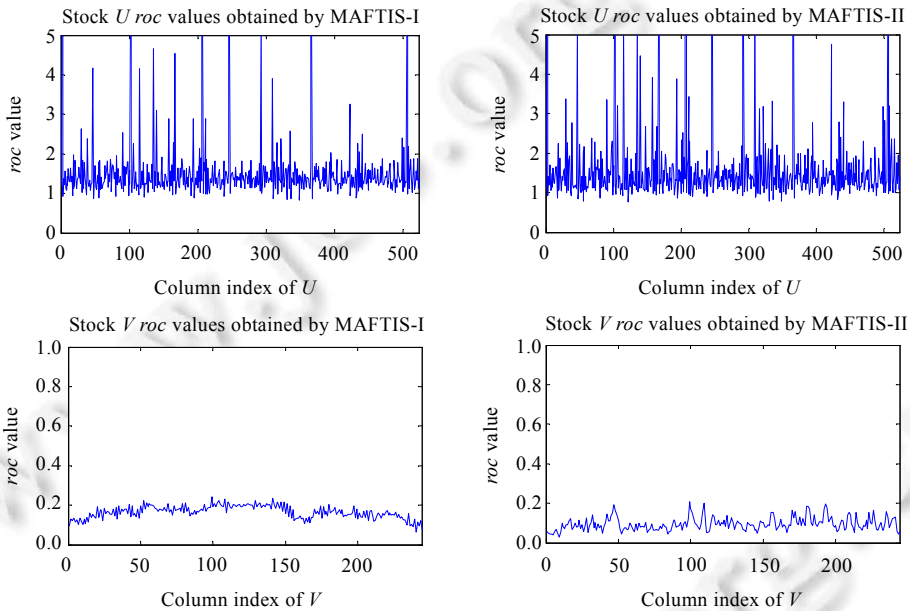

Fig.5 The *roc* values of the stock data

图 5 在 stock 数据上获得的 *roc* 值

综合图 5、图 6 的结果,矩阵 U 所对应的 *roc* 值在两个数据集的所有结果上的变化都没有明显的趋势;特别地,在相同的数据集上,由不同模型获得的 *roc* 值的分布也没有显著差别.这一现象支持了我们对 U 的解释,即:其所有的组成向量是独立同分布的,彼此的取值没有直接的影响.对于矩阵 V,我们发现:在 stock 数据集上,MAFTIS-I 对应的 *roc* 曲线比由 MAFTIS-II 得到的曲线更为光滑,显示在 MAFTIS-I 的解释下,影响股价的外部环境的变化趋势比由 MAFTIS-II 描述的结果更为稳定,因而 MAFTIS-I 的结果具有更好的可解释性.但我们同时注意到:MAFTIS-I 对应的曲线整体在 y=0.2 附近,MAFTIS-II 的曲线则围绕 y=0.1 作锯齿状的上下波动.而前文对 stock 数据的分析显示,其|*roi*|值主要聚集在以 y=0.1 为中心的区域,这与 MAFTIS-II 曲线的取值区间相吻合.这一现象也解释了在表 2 的结果中,MAFTIS-II 的表现优于 MAFTIS-I 的原因.

对于 traffic 数据集,由图 6 第 2 行的两图可以看到:由 MAFTIS-I 和 MAFTIS-II 得到的 V 矩阵的 *roc* 值变化都非常剧烈,然而 MAFTIS-I 对应的 *roc* 值主要分布在区间[0.1,0.6]中,MAFTIS-II 对应的 *roc* 值则主要落在区间[0.2,0.4]内.对比图 2 中 traffic 数据的 *roi* 值分布,由 MAFTIS-I 获得的结果与其更为接近.

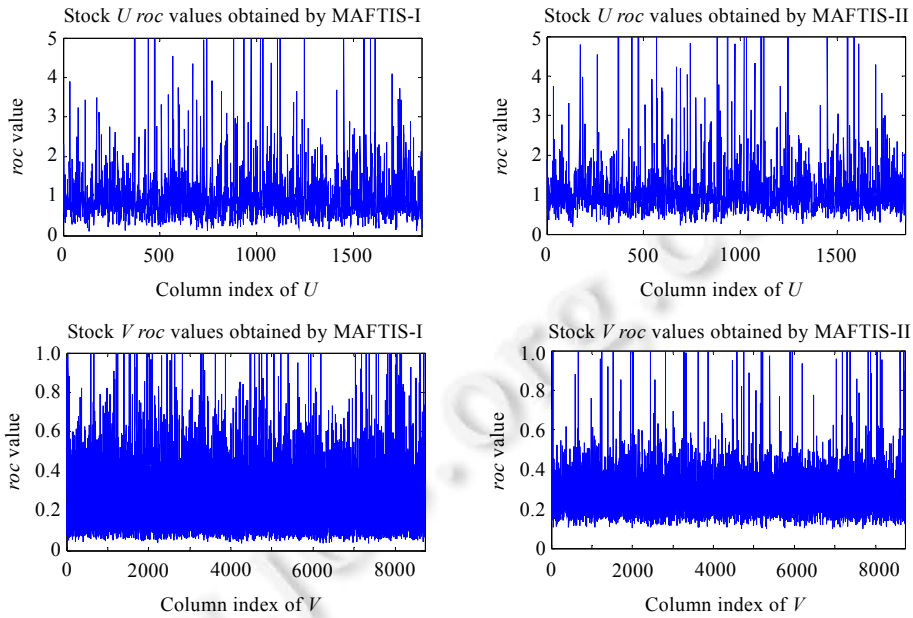

Fig.6 The *roc* values of the traffic data

图 6 在 traffic 数据上获得的 *roc* 值

此外,由于图 6 中的数据过于密集,为更好地了解其 *V* 矩阵的 *roc* 值变化的特点,我们在图 7 中以 0.2 为最大上界,分别统计了对于分别由 MAFTIS-I 和 MAFTIS-II 获得的 *V* 矩阵 *roc* 值的累积密度分布.在图 7 中,'*'对应 MAFTIS-I,'+'对应 MAFTIS-II.可以看出:对于 MAFTIS-I 获得的结果,有超过 50%的 *roc* 值在 0.2 以下;而对于 MAFTIS-II,这一比例仅不到 30%.两者比较,前者反映的变化趋势更为稳定.也支持了我们基于 stock 数据集的 *roc* 值分析所作的论断,即,MAFTIS-I 的拟合结果具有更好的可解释性.

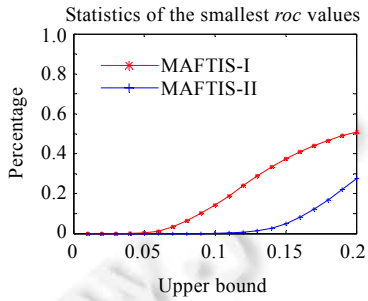

Fig.7 The cdf of the *roc* values of the *V* factor matrix obtained from the traffic data

图 7 Traffic 数据的 *V* 因子矩阵的 *roc* 值分布统计

- 收敛速度

同样使用前文 Given 50%的设置,分别记录了 MAFTIS-I,MAFTIS-II 在 stock 和 traffic 两个数据集上前 300 次循环中计算获得的 RMSE,具体如图 8 所示.

图 8 第 1 行对应了在 stock 数据上的结果,第 2 行对应了在 traffic 数据上的结果;左侧子图对应 MAFTIS-I,右侧子图对应 MAFTIS-II.所有子图的横坐标记录循环的次数,纵坐标代表 RMSE 的取值.

从图 8 可以看出,MAFTIS-I 和 MAFTIS-II 的收敛速度较为一致.在 stock 数据集上,它们的 RMSE 在前 100 次循环的下降比较明显,其后几乎保持平稳;在 traffic 数据集上,在前 50 个循环后,其 RMSE 即基本保持稳定.这

也显示了 MAFTIS 在实际应用中的有效性.

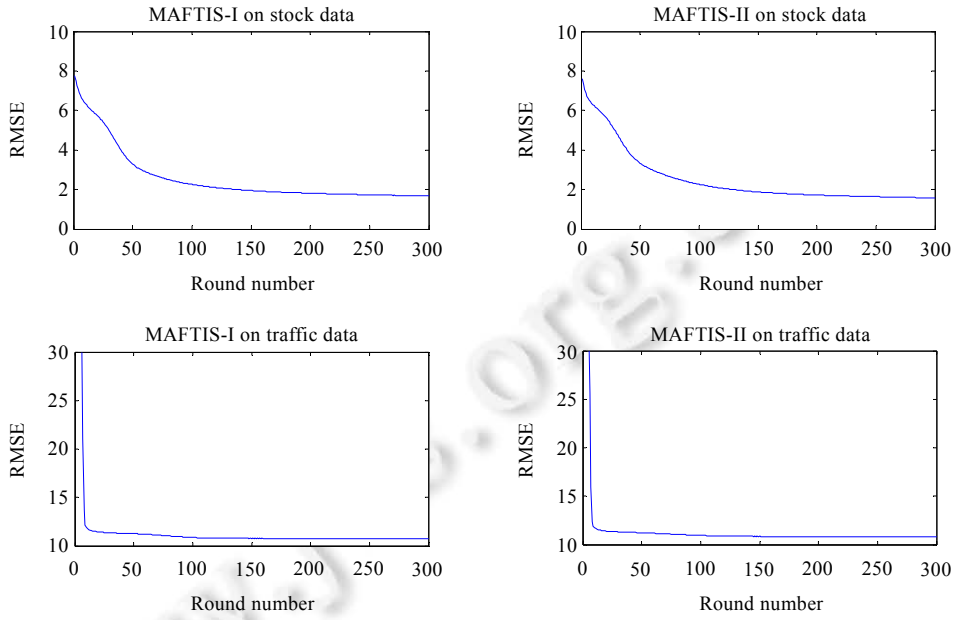

Fig.8 The convergence speed of MAFTIS on the stock and traffic data

图 8 MAFTIS 的收敛速度分析

5 总结与工作展望

时序数据是日常生活中最为常见的数据类型之一,在本文的工作中,我们提出了一种适用于时序数据的矩阵分解模型 MAFTIS:我们首先提出了一个适合于描述时序数据的概率图模型,在该模型中,我们提出使用两种不同的分布来描述时序数据中相邻时间点上的数据间的联系;根据这一描述,进而我们相应地导出了两种不同的矩阵分解模型 MAFTIS-I 和 MAFTIS-II.

我们分别给出了对 MAFTIS-I,MAFTIS-II 的求解策略,并在两个不同性质的时序数据集上检验了本文模型在数据恢复应用中的性能表现.我们的实验结果显示:当数据在时间轴上是以渐变的方式发生变化时,MAFTIS-II 可能会获得最好的预测结果;而当数据的变化是以突变的形式发生时,MAFTIS-I 可能更为合适.

容易看出,本文算法成功的关键在于结构化先验的引入.然而在另一方面,出于实际计算效率的考虑,本文只讨论了确定的链式先验的情况.

显然,本文工作的一个重要推广方向是对不确定的复杂先验的矩阵分解的研究.为此,我们至少还需解决如下两个子问题:

(1) 如何根据不完整的数据获得数据之间的结构关系?

(2) 如何把复杂的结构先验融入到矩阵的分解过程之中?

其中,对于问题(1),文献[30,31]等工作已经作了有益的探索,这将对问题(2)的研究起到一定的促进作用.然而,考虑到复杂先验对计算效率的影响,对适用于大规模数据的高效矩阵分解算法的研究仍将是我们未来关注的焦点之一.

References:

[1] Bell R, Koren Y, Volinsky C. Matrix factorization techniques for recommender systems. IEEE Computer, 2009. 42–49. [doi: 10.1109/MC.2009.263]

[2] Koren K. Factorization meets the neighborhood: A multifaceted collaborative filtering model. In: Proc. of the 14th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. ACM Press, 2008. 426−434. [doi: 10.1145/1401890.1401944]

[3] Salakhutdinov R, Mnih A. Probabilistic matrix factorization. In: Proc. of the NIPS. 2007.

[4] Xu MJ, Zhu J, Zhang B. Bayesian nonparametric maximum margin matrix factorization for collaborative prediction. In: Proc. of the Advances in Neural Information Processing Systems (NIPS 2012). 2012.

[5] Srebro N, Rennie JDM, Jaakkola T. Maximum margin matrix factorization. In: Proc. of the Advances in Neural Information Processing Systems 17.

[6] Balakrishnan S, Chopra S. Collaborative ranking. In: Proc. of the 5th Int'l Conf. on Web Search and Data Mining. 2012. [doi: 10.1145/2124295.2124314]

[7] Weimer M, Karatzoglou A, Le QV, Smola A. CofiRank—Maximum margin matrix factorization for collaborative ranking. In: Proc. of the NIPS. 2007.

[8] Krohn-Grimberghe A, Drumond L, Freudenthaler C. Multi-Relational matrix factorization using bayesian personalized ranking for social network data. In: Proc. of the 5th ACM Int'l Conf. on Web Search and Data Mining. 2012. 173−182. [doi: 10.1145/2124295.2124317]

[9] Vapnik VN. Statistical Learning Theory. Wiley-Interscience, 1998.

[10] Fazel M, Hindi H, Boyd S. A rank minimization heuristic with application to minimum order system approximation. In: Proc. of the American Control Conf. Arlington, 2001. [doi: 10.1109/ACC.2001.945730]

[11] Recht B, Fazel M, Parrilo P. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. SIAM Review, 2008,52(3):471−501. [doi: 10.1137/070697835]

[12] Candès EJ, Recht B. Exact matrix completion via convex optimization. Foundations of Computational Mathematics, 2008,9: 717−772. [doi: 10.1007/s10208-009-9045-5]

[13] Candès EJ, Tao T. The power of convex relaxation: Near-Optimal matrix completion. IEEE Trans. on Information Theory, 2009,56(5): 2053−2080. [doi: 10.1109/TIT.2010.2044061]

[14] Srebro N, Schraibman A. Rank, trace-norm and maxnorm. In: Proc. of the 18th Annual Conf. on Learning Theory. 2005. [doi: 10.1007/11503415_37]

[15] Koller D, Friedman N. Probabilistic Graphical Models: Principles and Techniques. MIT Press, 2009.

[16] Salakhutdinov R, Mnih A. Bayesian probabilistic matrix factorization using MCMC. In: Proc. of the ICML. 2008.

[17] Adams R, Dahl G, Murray I. Incorporating side information in probabilistic matrix factorization with gaussian processes. In: Proc. of the UAI. 2010.

[18] Porteous I, Asuncion A, Welling M. Bayesian matrix factorization with side information and dirichlet process mixtures. In: Proc. of the AAAI. 2010.

[19] Stephen B, Vandenberghe L. Convex Optimization. Cambridge University Press, 2009.

[20] Candes EJ, Romberg J, Tao T. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. IEEE Trans. on Information Theory, 2006,52(2):489−509. [doi: 10.1109/TIT.2005.862083]

[21] Bertsekas DP. Nonlinear Programming. 2nd ed., Athena Scientific, 1999.

[22] Tibshirani R, Saunders M. Sparsity and smoothness via the fused lasso. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 2005,67(1):91−108. [doi: 10.1111/j.1467-9868.2005.00490.x]

[23] Jerome F, Hastie T, Hofling H, Tibshirani R. Pathwise coordinate optimization. The Annals of Applied Statistics, 2007,1(2): 302−332. [doi: 10.1214/07-AOAS131]

[24] Golub, GH, Van Loan CF. Matrix Computations. Vol.3, JHU Press, 2012.

[25] Marlin B. Collaborative filtering: A machine learning perspective [MS. Thesis]. University of Toronto, Computer Science Department, 2004.

[26] Tipping ME, Bishop CM. Probabilistic principal component analysis. Journal of the Royal Statistical Society (Series B), 1999,61(3): 611−622. [doi: 10.1111/1467-9868.00196]

[27] Qu L, Hu JM, Li L, Zhang Y. PPCA-Based missing data imputation for traffic flow volume: A systematical approach. IEEE Trans. on Intelligent Transportation Systems, 2009. [doi: 10.1109/TITS.2009.2026312]

[28] http://lear.inrialpes.fr/people/verbeek/software.php

[29] http://www.utstat.toronto.edu/rsalakhu/BPMF.html

[30] Kolar M, Xing E. Consistent covariance selection from data with missing values. In: Proc. of the ICML. 2012.

[31] Stadler N, Buhlmann P. Missing values: Sparse inverse covariance estimation and an extension to sparse regression. Statistics and Computing, 2009. [doi: 10.1007/s11222-010-9219-7]

黄晓宇(1977－),男,广东茂名人,博士,讲师,主要研究领域为机器学习理论.

潘嵘(1976－),男,博士,副教授,主要研究领域为数据挖掘,机器学习.

李磊(1951－),男,博士,教授,博士生导师,主要研究领域为人工智能理论.

梁冰(1978－),女,工程师,主要研究领域为位置服务技术,推荐系统.

陈康(1972－),男,工程师,主要研究领域为位置服务技术,推荐系统.

蔡文学(1968－),男,博士,教授,主要研究领域为位置服务技术,智能交通系统.

www.jos.org.cn