浙江理工大学学报(自然科学版),第 37 卷,第 6 期,2017 年 11 月
Journal of Zhejiang Sci-Tech University (Natural Sciences)
Vol. 37, No. 6, Nov. 2017

DOI:10.3969/j.issn.1673-3851.2017.11.020

矩阵填充算法在抗癌药物敏感性研究中的运用

黄　莉,贺平安

(浙江理工大学理学院,杭州 310018)

摘　要:从不完整的数据推断完整有效的数据,继而对原始数据给出可靠的分析是一个重要的数学问题。根据低秩矩阵填充算法,提出一种融合癌细胞系基因表达数据相似性信息的低秩矩阵填充算法。应用该算法对癌细胞系与抗癌药物反应的敏感性缺失数据进行恢复,并对相对反应低的数值进行重评估。利用均方根误差和 10 倍交叉验证法评估该算法,结果显示该算法比已有算法的均方根误差减少 22.7%,说明该算法具有很好的数据恢复效果。

关键词:抗癌药物敏感性;低秩矩阵填充;癌细胞系;10 倍交叉验证;均方根误差

中图分类号:O29　　**文献标志码**:A　　**文章编号**:1673-3851 (2017) 06-0881-07

0　引　言

肿瘤是在体内外各种因素的作用下由一系列基因连续突变导致细胞生长失去控制所致,因而对于每个肿瘤患者,即使是同一种肿瘤,其致病因素和体内突变的基因可能都不相同。每一个患者的肿瘤都有独特的生物特征,即肿瘤的异质性[1]。肿瘤异质性要求对不同的肿瘤患者要不同对待,即肿瘤的个性化医疗[2],目前,以患者个人遗传信息为基础的个性化医疗已成为医学领域主要研究方向之一。

在临床试验中,为了研究和获得针对于每个肿瘤患者有效个性化医疗,通常采用漫长而昂贵的药物开发与实验验证来评估药物的疗效和毒性,但资源的稀缺限制该方案的实际应用。而解决该问题可能性方案是将肿瘤细胞在体外进行二维或三维培养,再直接测定病人肿瘤细胞的药物敏感性[3],或者采用癌细胞体内培养模型,如小鼠异种移植模型和转基因小鼠模型[4]。该方案可以捕获大部分病人的肿瘤相关生物学特性;然而,这种方案代价大且耗时,缺乏可扩展性,测试药物数量最多只能到几百种。

在过去几十年中,随着高通量技术发展,研究人员提出一种替代方案,即利用癌细胞系大面板建立药物反应的基因组预测因子[5-9]。目前预测药物敏感性的方法主要根据已知癌细胞株对不同药物的反应特性和功能基因组特征,对药物反应数据进行回归或分类[10]。

矩阵填充算法是根据已有的数据对缺失数据进行预测和恢复,Keshavan 等[11]对该算法的原理和正确性作了全面阐述和证明,Cai 等[12]在此基础上提出改进算法,并将该算法运用到 1968—2003 年 H3N2 数据上,对血凝素抑制试验中缺失数据进行预测。

本文根据低秩矩阵填充算法,提出一种新型预测药物敏感性的方法,将癌细胞系基因表达数据相似性信息融合进已有低秩矩阵填充算法中。利用该算法对癌细胞系百科全书(cancer cell line encyclopedia, CCLE)中 491 种癌细胞系与 24 种抗癌药物反应的敏感性数据进行缺失数据填充和低反应数据重估,使得 CCLE 数据库中抗癌药物反应的敏感性数据更加详尽和完整;获得模型中参数 r、λ_1、λ_2 的最优取值,并通过与已有算法比较证明该算法的有效性。

收稿日期:2017－03－03　　网络出版日期:2017－05－24

基金项目:国家自然科学基金项目(61170110,61272312);浙江省自然科学基金项目(LY14F020049)

作者简介:黄　莉(1990－),女,湖北黄冈人,硕士研究生,主要从事生物信息学方面的研究。

通信作者:贺平安,E-mail:pinganhe@zstu.edu.cn

1 数 据

1.1 抗癌药物反应敏感性数据

CCLE 是由哈佛大学、麻省理工 Broad 研究院和诺华生物研究所等研究机构开发的首个可免费获取的癌症基因组数据大型公共资源平台[13]。CCLE 整合来自大约 1036 种人类癌症细胞系的基因表达、染色体拷贝数等大规模数据，并且还包含 504 种细胞系与 24 种抗癌药物反应的敏感性数据，并覆盖 36 种常见癌症类型[14]。CCLE 旨在为癌症研究提供数据支持以获得更多的发现，通过理解癌症细胞系与药物之间关系获取抗癌药物潜在的敏感性相关标志物，最终为癌症治疗寻找合适药物[15]。

本文从 CCLE 数据库(http://www.broadinstitute.org/ccle)下载 504 种癌细胞系与 24 种药物反应的药物敏感性数据及 1036 种癌细胞系的基因表达数据。并将 504 种癌细胞系与 1036 种癌细胞系进行匹配，发现其中只有 491 种癌细胞系相同。因此，本文实际运用数据为 491 种癌细胞系与 24 种药物反应的药物敏感性数据以及这 491 种癌细胞系的基因表达数据。表 1 为部分抗癌药物反应敏感性数据。

表 1 抗癌药物反应敏感性数据

癌细胞系	抗癌药物					
	L-685458	ZD-6474	Sorafenib	Irinotecan	PD-0325901	PD-0332991
BT474_BREAST	0.7397	0.5179	0.8083	NA	0.5865	1.0329
BT549_BREAST	0.1532	0.9727	0.5889	3.4086	0.1503	0
BXPC3_PANCREAS	0.3731	1.2382	0.3544	NA	2.8579	NA
C32_SKIN	0.4038×10^{-1}	0.4507	0.3158	2.0384	3.2117	0.7469×10^{-1}
C3A_LIVER	0.5550	1.4222	0.7144	NA	2.4419	0.4059
CAKI2_KIDNEY	0.1271	1.2657	0.4486	0.5537	1.8386	0.1423
CAL12T_LUNG	NA	2.5282	0.6577	NA	2.2204	NA
CAL27_UPPER	0	3.4375	0.3294	NA	2.6201	0
CAL78_BONE	0.7571	0.3050	0.3010	2.9043	0.1878	0.1654
CALU3_LUNG	0.4371	1.8192	0.5152	NA	1.8058	0.3678
CALU6_LUNG	NA	0.7187	0.8441	NA	3.2483	NA
CAPAN2_PANCREAS	0	1.9235	0	1.5901	3.1150	0.3713

注：表中“NA”表示敏感值缺失。

表 1 数据表明，部分药物敏感性数据缺失，此外部分药物敏感性数据值为 0，因此，药物敏感性数据大致可以被分为 3 种类型：第一类数据，大于 0 的敏感值；第二类数据，等于 0 的敏感值；第三类数据，敏感值缺失。

1.2 数据预处理

为了方便分析，本文将 491 种癌细胞系与 24 种药物反应的药物敏感性数据简称 CD-491 数据集，其中的可观测数据有 11360 个。假设 m_{ij} 表示第 i 种癌细胞系与第 j 种药物反应敏感值，需要对原始的药物敏感性数据进行预处理，预处理包括两步：

第一步，标准化(normalizing)，对第二类和第三类数据进行预处理，具体操作如下，

$$m'_{ij}=\begin{cases} m_{ij}, & m_{ij}\ \text{属于第一类型数据} \\ \dfrac{\min(m_{ij})}{100}, & m_{ij}\ \text{属于第二类型数据.} \\ 0, & m_{ij}\ \text{属于第三类型数据} \end{cases}$$

其中，$\min(m_{ij})$ 表示 CD-491 数据集中非零最小值。

第二步，修整(trimming)，为了避免奇异向量高度集中在高权重的列(或行)，需要将矩阵中一些观察值随机设为 0[11]。对于 CD-491 数据集中每一行，如果观察值个数大于 $|\boldsymbol{E}|/m$($|\boldsymbol{E}|$ 表示矩阵 $\boldsymbol{E}$ 中观察值个数，m 表示矩阵 $\boldsymbol{E}$ 的行数)，就随机设置一些观察值为 0。同样地，对于每一列，如果观察值个数大于 $|\boldsymbol{E}|/n$(n 表示矩阵 $\boldsymbol{E}$ 的列数)，随机将一些观察值设置为 0。

2 方法和结果

2.1 低秩矩阵填充模型

矩阵填充就是对矩阵中缺失的数据进行恢复和已有数据的矫正。假设 $\boldsymbol{M}=(m_{ij})_{m\times n}$ 的矩阵和一个集合 $\boldsymbol{E}$，矩阵 $\boldsymbol{M}$ 的行表示是 m 种癌细胞系，列表示 n 种抗癌药物，m_{ij} 表示第 i 种癌细胞系与第 j 种药物反应的敏感值，反应值可以分为 3 种类型，而 $\boldsymbol{E}$ 为第

一类型数据与第二类型数据位置的集合,即$(i,j)\in \boldsymbol{E}\subseteq[m]\times[n]$。假设矩阵$\boldsymbol{M}$是秩为$r$的低秩矩阵,且$r\ll m,n$,则由奇异值分解[16-17]可知:存在矩阵$\boldsymbol{U}_{m\times r}$、矩阵$\boldsymbol{V}_{n\times r}$和矩阵$\boldsymbol{\Sigma}_{r\times r}$,使得$\boldsymbol{M}=\boldsymbol{U}_{m\times r}\boldsymbol{\Sigma}_{r\times r}(\boldsymbol{V}_{n\times r})^{\mathrm{T}}$。

标准的矩阵填充模型如下:

$$\min \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{n}(\boldsymbol{X}_{ij}^{\boldsymbol{E}}-\boldsymbol{M}_{ij}^{\boldsymbol{E}})^2+\lambda_1 G(\boldsymbol{X})$$
$$\text{s.t.}\quad \boldsymbol{X}=\boldsymbol{U}_{m\times r}\boldsymbol{\Sigma}_{r\times r}(\boldsymbol{V}_{n\times r})^{\mathrm{T}} \tag{1}$$

其中:$\boldsymbol{X}=\boldsymbol{U}_{m\times r}\boldsymbol{\Sigma}_{r\times r}(\boldsymbol{V}_{n\times r})^{\mathrm{T}}$是被估计的矩阵,$\boldsymbol{X}_i$、$\boldsymbol{X}_j$分别表示矩阵$\boldsymbol{X}$的第$i$行和第$j$行。$\boldsymbol{M}^{\boldsymbol{E}}$表示观察矩阵,且满足条件$\boldsymbol{M}_{ij}^{\boldsymbol{E}}=\begin{cases}m_{ij}, & (i,j)\in \boldsymbol{E}\\ 0, & \text{其他情况}\end{cases}$。函数$G(\boldsymbol{X})$是一个正则化项,本文取$G(\boldsymbol{X})=\sum_{i=1}^{m}g\left(\frac{\|\boldsymbol{U}^i\|^2}{3\alpha r}\right)+\sum_{i=1}^{n}g\left(\frac{\|\boldsymbol{V}^i\|^2}{3\alpha r}\right)$,且当$Z\geqslant 1$时,$g(z)=\mathrm{e}^{(z-1)^2}-1$,否则$g(Z)=0$;$\boldsymbol{U}^i$、$\boldsymbol{V}^i$分别表示$\boldsymbol{U}$和$\boldsymbol{V}$的第$i$行;$\alpha=\max\{m,n\}$。

Cai 等[12]在构建流感病毒血凝素抑制试验数据恢复模型时认为,上述矩阵填充模型只是充分运用第一类型数据信息,虽然第二类型数据比第一类数据信息少,但能提高矩阵填充后数据准确性,于是,引入一个阈值θ_{ij},当矩阵中数据属于第二类型数据时,θ_{ij}被设置成常数C;若该数据不属于第二类型数据,$\theta_{ij}=-\infty$。在模型(1)基础上,提出一个修正的融合第二类数据信息的低秩矩阵填充模型:

$$\min \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{n}(\boldsymbol{X}_{ij}^{\boldsymbol{E}}-\boldsymbol{M}_{ij}^{\boldsymbol{E}})^2 I(\boldsymbol{X}_{ij}^{\boldsymbol{E}}\geqslant\theta_{ij})+\lambda_1 G(\boldsymbol{X})$$
$$\text{s.t.}\quad \boldsymbol{X}=\boldsymbol{U}_{m\times r}\boldsymbol{\Sigma}_{r\times r}(\boldsymbol{V}_{n\times r})^{\mathrm{T}} \tag{2}$$

当$\boldsymbol{X}_{ij}^{\boldsymbol{E}}\geqslant\theta_{ij}$时,$I(\boldsymbol{X}_{ij}^{\boldsymbol{E}}\geqslant\theta_{ij})=1$,否则$I(\boldsymbol{X}_{ij}^{\boldsymbol{E}}\geqslant\theta_{ij})=0$。该模型第一项是为了减少观察值与估计值之间的误差,第二项是对被估计矩阵的正则化。

但是,上述两个模型都仅仅是从数据的本身出发,忽略数据中对象的本身特征信息学,例如蛋白质序列相似性信息,基因表达信息等。本文在模型(2)基础上提出一种融合数据对象的特征信息的低秩矩阵填充模型:

$$\min \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{n}(\boldsymbol{X}_{ij}^{\boldsymbol{E}}-\boldsymbol{M}_{ij}^{\boldsymbol{E}})^2 I(\boldsymbol{X}_{ij}^{\boldsymbol{E}}\geqslant\theta_{ij})+$$
$$\lambda_1 G(\boldsymbol{X})+\lambda_2\sum_{i=1}^{m-1}\sum_{j=i+1}^{m}s_{ij}\|\boldsymbol{X}_i-\boldsymbol{X}_j\|^2$$
$$\text{s.t.}\quad \boldsymbol{X}=\boldsymbol{U}_{m\times r}\boldsymbol{\Sigma}_{r\times r}(\boldsymbol{V}_{n\times r})^{\mathrm{T}} \tag{3}$$

其中:s_{ij}表示癌细胞系i与癌细胞系j基因表达数据的Person相关性系数,$\boldsymbol{X}_i$表示矩阵$\boldsymbol{X}$的第i行。此模型是基于以下观点提出的:如果两个癌细胞系的基因表达数据相关性系数越高,那么它们的药物反应敏感性数据越相似。即当这些癌细胞系被投射到一个几何空间中,相关性系数越高的癌细胞系,它们的空间距离越近。

2.2　模型算法

为了求解模型(3),本文提出了一种基于梯度下降的算法,算法迭代步骤如下:

步骤 1　对$\boldsymbol{M}^{\boldsymbol{E}}$进行奇异值分解,即$\boldsymbol{M}^{\boldsymbol{E}}=\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^{\mathrm{T}}$,分别设置$\boldsymbol{U}^{(0)}$和$\boldsymbol{V}^{(0)}$的初始为$\boldsymbol{U}^{(0)}=\boldsymbol{U}_0*\sqrt{m}$,$\boldsymbol{V}^{(0)}=\boldsymbol{V}_0*\sqrt{n}$,$\boldsymbol{U}_0$和$\boldsymbol{V}_0$分别是取$\boldsymbol{U}$和$\boldsymbol{V}$的前$r$列的两个矩阵;

步骤 2　固定$\boldsymbol{U}^{(k)}$和$\boldsymbol{V}^{(k)}$,通过最小化$\Sigma\Sigma(\boldsymbol{X}_{ij}^{\boldsymbol{E}}-\boldsymbol{M}_{ij}^{\boldsymbol{E}})^2 I(\boldsymbol{X}_{ij}^{\boldsymbol{E}}\geqslant\theta_{ij})$计算相应的$\boldsymbol{\Sigma}_{r\times r}$;

步骤 3　通过梯度下降法更新$\boldsymbol{U}^{(k+1)}$和$\boldsymbol{V}^{(k+1)}$:即$\boldsymbol{U}^{(k+1)}=\boldsymbol{U}^{(k)}+t\cdot\nabla\boldsymbol{U}^{(k)}$,$\boldsymbol{V}^{(k+1)}=\boldsymbol{V}^{(k)}+t\cdot\nabla\boldsymbol{V}^{(k)}$,$\nabla\boldsymbol{U}$和$\nabla\boldsymbol{V}$分别表示$\boldsymbol{U}$和$\boldsymbol{V}$的梯度;

步骤 4　重复步骤 2 和步骤 3,当该算法收敛(迭代误差小于10^{-8})或者达到某一给定的迭代次数(本文设置 2000 次)停止。

上述模型算法中需要分别计算$\boldsymbol{U}$和$\boldsymbol{V}$的梯度。由于本文模型比 Cai 等[12]提出的模型(2)增加一项:

$$\lambda_2\sum_{i=1}^{m-1}\sum_{j=i+1}^{m}s_{ij}\|X_i-X_j\|^2 \tag{4}$$

因此,需要对 Cai 等[12]的算法加以改进。为了求解模型中$\boldsymbol{U}$和$\boldsymbol{V}$的梯度,本文首先引入引理 1。

引理 1[18]　设$\boldsymbol{A}=(a_{ij})\in\mathbf{R}^{m\times m}$和$\boldsymbol{B}=(b_{ij})\in\mathbf{R}^{m\times m}$都是常数矩阵,$\boldsymbol{X}=(x_{ij})\in\mathbf{R}^{m\times m}$是一个变量矩阵,则有

$$\frac{\partial tr(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{X}\boldsymbol{B})}{\partial X}=\boldsymbol{A}\boldsymbol{X}\boldsymbol{B}+\boldsymbol{A}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{B}^{\mathrm{T}} \tag{5}$$

$$\frac{\partial tr(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{A})}{\partial \boldsymbol{X}}=\boldsymbol{X}\boldsymbol{A}+\boldsymbol{X}\boldsymbol{A}^{\mathrm{T}} \tag{6}$$

不妨令

$$h(\boldsymbol{X})=\lambda_2\sum\sum s_{ij}\|\boldsymbol{X}_i-\boldsymbol{X}_j\|^2,$$

则有

$$\begin{aligned}h(\boldsymbol{X})&=\lambda_2\sum\sum s_{ij}\|\boldsymbol{X}_i-\boldsymbol{X}_j\|^2\\&=2\lambda_2\sum_{i>j}s_{ij}\|\boldsymbol{X}_i-\boldsymbol{X}_j\|^2\\&=2\lambda_2\,\mathrm{tr}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{X}),\end{aligned}$$

其中:$\boldsymbol{X}=\boldsymbol{U}_{m\times r}\boldsymbol{\Sigma}_{r\times r}(\boldsymbol{V}_{n\times r})^{\mathrm{T}}$,$\boldsymbol{A}=(a_{ij})_{m\times m}$是一个对称矩阵,且$a_{ij}=\begin{cases}\sum_{k\neq i}s_{i\neq k}, & \text{当 } i=j\\ -s_{ij}, & \text{当 } i\neq j\end{cases}$。

根据迹的性质,有

$$\begin{aligned}\mathrm{tr}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{X}) &= \mathrm{tr}((\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^{\mathrm{T}})^{\mathrm{T}}\boldsymbol{A}\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^{\mathrm{T}})\\ &= \mathrm{tr}(\boldsymbol{V}\boldsymbol{\Sigma}^{\mathrm{T}}\boldsymbol{U}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^{\mathrm{T}})\\ &= \mathrm{tr}(\boldsymbol{U}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^{\mathrm{T}}\boldsymbol{V}\boldsymbol{\Sigma}^{\mathrm{T}})\\ &= \mathrm{tr}(\boldsymbol{V}^{\mathrm{T}}\boldsymbol{V}\boldsymbol{\Sigma}^{\mathrm{T}}\boldsymbol{U}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{U}\boldsymbol{\Sigma}).\end{aligned}$$

由式(5)可知：

$$\begin{aligned}\frac{\partial h(\boldsymbol{X})}{\partial \boldsymbol{U}} &= 2\lambda_2\frac{\partial}{\partial \boldsymbol{U}}\mathrm{tr}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{X})\\ &= 2\lambda_2\frac{\partial}{\partial \boldsymbol{U}}\mathrm{tr}(\boldsymbol{U}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^{\mathrm{T}}\boldsymbol{V}\boldsymbol{\Sigma}^{\mathrm{T}})\\ &= 2\lambda_2(\boldsymbol{A}\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^{\mathrm{T}}\boldsymbol{V}\boldsymbol{\Sigma}^{\mathrm{T}}+\boldsymbol{A}^{\mathrm{T}}\boldsymbol{U}(\boldsymbol{\Sigma}\boldsymbol{V}^{\mathrm{T}}\boldsymbol{V}\boldsymbol{\Sigma}^{\mathrm{T}})^{\mathrm{T}})\\ &= 4\lambda_2\boldsymbol{A}\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^{\mathrm{T}}\boldsymbol{V}\boldsymbol{\Sigma}^{\mathrm{T}}.\end{aligned}$$

由式(6)可知：

$$\begin{aligned}\frac{\partial h(\boldsymbol{X})}{\partial \boldsymbol{V}} &= 2\lambda_2\frac{\partial}{\partial \boldsymbol{V}}\mathrm{tr}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{X})\\ &= 2\lambda_2\frac{\partial}{\partial \boldsymbol{U}}\mathrm{tr}(\boldsymbol{V}^{\mathrm{T}}\boldsymbol{V}\boldsymbol{\Sigma}^{\mathrm{T}}\boldsymbol{U}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{U}\boldsymbol{\Sigma})\\ &= 2\lambda_2(\boldsymbol{V}\boldsymbol{\Sigma}^{\mathrm{T}}\boldsymbol{U}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{U}\boldsymbol{\Sigma}+\boldsymbol{V}(\boldsymbol{\Sigma}^{\mathrm{T}}\boldsymbol{U}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{U}\boldsymbol{\Sigma})^{\mathrm{T}})\\ &= 4\lambda_2\boldsymbol{V}\boldsymbol{\Sigma}^{\mathrm{T}}\boldsymbol{U}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{U}\boldsymbol{\Sigma}.\end{aligned}$$

结合 Cai 等[12] 模型前两项 $\boldsymbol{U}$ 和 $\boldsymbol{V}$ 的梯度，可以得出整个模型中 $\boldsymbol{U}$ 和 $\boldsymbol{V}$ 的梯度分别是：

$$\nabla\boldsymbol{U} = ((\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^{\mathrm{T}})^{\boldsymbol{E}} - \boldsymbol{M}^{\boldsymbol{E}})\cdot\boldsymbol{I}\boldsymbol{V}\boldsymbol{\Sigma}^{\mathrm{T}} + \boldsymbol{U}\boldsymbol{Q}_U + \lambda_1 f(\boldsymbol{U}, 2\mathrm{e}^{(\boldsymbol{Q}_{U1}-\boldsymbol{I}_1)^2}(\boldsymbol{Q}_{U1}-\boldsymbol{I}_1)) + 4\lambda_2\boldsymbol{A}\boldsymbol{U}\boldsymbol{\Sigma}(\boldsymbol{V}^{\mathrm{T}}\boldsymbol{V})\boldsymbol{\Sigma}^{\mathrm{T}},$$

$$\nabla\boldsymbol{V} = ((\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^{\mathrm{T}})^{\boldsymbol{E}} - \boldsymbol{M}^{\boldsymbol{E}})^{\mathrm{T}}\cdot\boldsymbol{I}\boldsymbol{U}\boldsymbol{\Sigma} + \boldsymbol{V}\boldsymbol{Q}_V + \lambda_1 f(\boldsymbol{V}, 2\mathrm{e}^{(\boldsymbol{Q}_{v1}-\boldsymbol{I}_2)^2}(\boldsymbol{Q}_{v1}-\boldsymbol{I}_2)) + 4\lambda_2\boldsymbol{V}\boldsymbol{\Sigma}^{\mathrm{T}}\boldsymbol{U}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{U}\boldsymbol{\Sigma}.$$

其中：$\boldsymbol{I}$ 是一个指示矩阵，当 $\boldsymbol{X}_{ij}^{\boldsymbol{E}} \geqslant \theta_{ij}$ 时，$\boldsymbol{I}_{ij}=1$，否则，$\boldsymbol{I}_{ij}=0$；“·”表示矩阵中对应元素相乘；

$$\boldsymbol{Q}_U = \frac{1}{m}\boldsymbol{U}^{\mathrm{T}}(\boldsymbol{M}^{\boldsymbol{E}} - (\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^{\mathrm{T}})^{\boldsymbol{E}})\boldsymbol{V}\boldsymbol{\Sigma}^{\mathrm{T}},$$

$$\boldsymbol{Q}_V = \frac{1}{n}\boldsymbol{V}^{\mathrm{T}}(\boldsymbol{M}^{\boldsymbol{E}} - (\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^{\mathrm{T}})^{\boldsymbol{E}})^{\mathrm{T}}\boldsymbol{U}\boldsymbol{\Sigma},$$

$\boldsymbol{I}_1=(1,1,\cdots,1)_{m\times1}^{\mathrm{T}}\boldsymbol{I}_2=(1,1,\cdots,1)_{n\times1}^{\mathrm{T}}$，$\alpha=\max\{m,n\}$，

$$\boldsymbol{Q}_{u1}=\frac{1}{3\alpha r}\begin{bmatrix}\sum_{j=1}^{r}\boldsymbol{U}_{1j}^2\\ \sum_{j=1}^{r}\boldsymbol{U}_{2j}^2\\ \vdots\\ \sum_{j=1}^{r}\boldsymbol{U}_{mj}^2\end{bmatrix},\boldsymbol{Q}_{v1}=\frac{1}{3\alpha r}\begin{bmatrix}\sum_{j=1}^{r}\boldsymbol{V}_{1j}^2\\ \sum_{j=1}^{r}\boldsymbol{V}_{2j}^2\\ \vdots\\ \sum_{j=1}^{r}\boldsymbol{V}_{nj}^2\end{bmatrix},$$

$$f(\boldsymbol{A}_{m\times r},\boldsymbol{B}_{m\times1})=\boldsymbol{Z}_{m\times r},\boldsymbol{Z}_{ij}=\begin{cases}\frac{1}{\alpha r}\boldsymbol{A}_{ij}*\boldsymbol{B}_i, & \text{当 } B_i>0\\ 0, & \text{其它}\end{cases}。$$

2.3 算法评估

评价一个模型的好坏，有许多参数标准，比如和方差（*SSE*）、均方根误差（*RMSE*）、确定系数（*R*-square）等。本文利用均方根误差来评估低秩矩阵填充模型的优劣。一般来说，均方根误差值越小，表示预测值越接近真实值[19]。

假设有两个向量 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$，且 $\boldsymbol{X}=(x_1,x_2,\cdots,x_k)$，$\boldsymbol{Y}=(y_1,y_2,\cdots,y_k)$，$x_i$、$y_i$ 分别代表观察值和相应的估计值，那么将 *RMSE* 定义为：

$$RMSE=\sqrt{\frac{\sum_{i=1}^{k}(x_i-y_i)^2}{n}}.$$

2.4 交叉验证

在本文矩阵填充算法中，利用 10 倍交叉验证法，通过训练参数 λ_1，λ_2，从而获取最小的 *RMSE* 值。所谓的 10 倍交叉验证法，就是均分样本数据为 10 组，选取其中的一组数据作为测试集，其余 9 组数据作为训练集，重复 10 次，将 10 次结果均值作为对算法精度估计的依据[20]。

本文将矩阵中已有的数据均分为 10 等份，每一次，选取其中的 9 份数据作为观察值进行矩阵的填充，然后计算填充完后的矩阵与观察矩阵在剩下 1 份数据集上的 *RMSE*，重复 10 次，将 10 次结果均值作为对该次算法精度的估计值，即为需要 *RMSE* 值。每次仅对测试集中第一类型数据计算 *RMSE*，*RMSE* 值也被称作局部 *RMSE* 值。

2.5 CD-491 数据集中缺失值的填充

首先将模型(2)运用到 CD-491 数据集中，用以预测数据集中的缺失值和重估值为 0 的数据。表 2 给出 10 倍交叉验证时，r，λ_1 取不同值情况下的 *RMSE*。从表 2 可以看出，当参数 $r=3$，$\lambda_1=1\times10^{-3}$ 时，模型(2)算法的 *RMSE* 最小，其值为 0.8654。

表 2 r 和 λ_1 取不同值时的 *RMSE*

r	λ_1						
	1×10^{-2}	1×10^{-3}	1×10^{-4}	1×10^{-5}	1×10^{-6}	1×10^{-7}	1×10^{-9}
2	1.2499	1.1725	1.1734	1.1248	1.1054	1.1223	1.2041
3	0.9024	0.8654	0.8678	0.8689	0.8758	0.9652	0.9674
4	0.9033	0.9053	0.9121	0.9135	0.9254	0.9285	0.9304
5	1.0241	1.0324	1.0344	0.9822	0.9853	0.9809	0.9818
6	1.1324	1.1236	1.1255	1.1276	1.1208	1.1019	1.1051
7	1.2320	1.2301	1.2322	1.2352	1.2321	1.2318	1.2318

进一步地,将模型(3) 运用到 CD-491 数据集中,为了得到完整矩阵,需要对 3 个未知参数进行训练,除了矩阵 $\boldsymbol{X}$ 的秩 r,模型中第二项系数 λ_1,还有第三项系数 λ_2,通过 10 倍交叉验证,分别获取参数 r、λ_1、λ_2 最优取值 $r=3$,$\lambda_1=1\times10^{-4}$,$\lambda_2=1\times10^{-9}$,此时 *RMSE* 值达到最小 0.6688,表 3－6 展示 r、λ_1、λ_2 取不同值情况下的 *RMSE*。

表 3　$r=3$ 时,λ_1 和 λ_2 取不同值时的 *RMSE*

λ_2	λ_1					
	1×10^{-2}	1×10^{-4}	1×10^{-5}	1×10^{-7}	1×10^{-9}	1×10^{-15}
1×10^{-2}	0.7463	0.6988	0.6725	0.6829	0.7452	0.8722
1×10^{-3}	0.8251	0.8054	0.7723	0.7525	0.7004	0.8421
1×10^{-4}	0.8085	0.7966	0.7658	0.7438	0.6688	0.8411
1×10^{-6}	0.8203	0.8024	0.8088	0.7695	0.7392	0.8776
1×10^{-8}	0.7587	0.6700	0.6700	0.6700	0.6700	0.6700
1×10^{-11}	0.7587	0.6699	0.6700	0.6700	0.6700	0.6700

表 4　$r=4$ 时,λ_1 和 λ_2 取不同值时的 *RMSE*

λ_2	λ_1					
	1×10^{-2}	1×10^{-4}	1×10^{-5}	1×10^{-7}	1×10^{-9}	1×10^{-15}
1×10^{-2}	0.7781	0.7688	0.7691	0.7709	0.7459	0.8261
1×10^{-3}	0.8521	0.8285	0.7536	0.7492	0.7324	0.8476
1×10^{-4}	0.7201	0.7388	0.7148	0.7110	0.7086	0.7083
1×10^{-6}	0.7262	0.7255	0.7238	0.7351	0.7473	0.8541
1×10^{-8}	0.7677	0.7800	0.7800	0.7800	0.7800	0.7800
1×10^{-11}	0.7800	0.7800	0.7800	0.7800	0.7800	0.7800

表 5　$r=5$ 时,λ_1 和 λ_2 取不同值时的 *RMSE*

λ_2	λ_1					
	1×10^{-2}	1×10^{-4}	1×10^{-5}	1×10^{-7}	1×10^{-9}	1×10^{-15}
1×10^{-2}	1.0049	0.9992	0.9960	0.9613	0.9614	0.9611
1×10^{-3}	0.8408	0.8405	0.8453	0.8524	0.8548	0.8521
1×10^{-4}	0.9544	0.9536	0.9506	0.9510	0.9254	0.9255
1×10^{-6}	1.0025	0.9917	0.9903	0.9723	0.9537	0.9255
1×10^{-8}	1.1256	1.1256	1.1256	1.1256	1.1256	1.1256
1×10^{-11}	1.1256	1.1256	1.1256	1.1256	1.1256	1.1256

表 6　$r=6$ 时,λ_1 和 λ_2 取不同值时的 *RMSE*

λ_2	λ_1					
	1×10^{-2}	1×10^{-4}	1×10^{-5}	1×10^{-7}	1×10^{-9}	1×10^{-15}
1×10^{-2}	1.1242	1.5685	1.1814	1.1762	1.1732	1.1743
1×10^{-3}	1.1757	1.1642	1.1504	1.1098	1.1082	1.1066
1×10^{-4}	1.0988	1.0988	1.0985	1.0910	1.0548	1.0529
1×10^{-6}	0.9865	0.9855	0.9846	0.9824	0.9827	0.9871
1×10^{-8}	1.0723	1.0721	1.0723	1.0723	1.0723	1.0723
1×10^{-11}	1.0723	1.0723	1.0723	1.0723	1.0723	1.0723

结合这3个参数的最优取值,利用模型(3),得到一个比较完整的CD-491数据集,原先由于受到试验条件的限制和干扰而无法直接获取的数据通过本文模型可以进行预测。此外,通过比较可以发现,模型(4)由于增加第三项,均方根误差由0.8654降到0.6688,即均方根误差减少22.7%,这说明融合癌细胞株的基因表达数据有利于矩阵填充。

3 结 论

在抗癌药物敏感性数据获取实验中,实验条件限制和外界因素的干扰会导致部分数据不准或部分数据缺失,利用数学方法从不完整数据中推断出可靠与完整抗癌药物敏感性数据是一个非常有意义的研究。本文根据低秩矩阵填充算法,提出一种合理、更具有生物意义低秩矩阵填充模型,该模型不仅仅从实验数据出发,而且充分利用数据对象本身的生物学特征,通过将癌细胞系基因表达数据融合进低秩矩阵填充模型,提高矩阵填充的准确性。

参考文献:

[1] 吴冠青,孙燕.恶性肿瘤的个性化治疗[J].癌症进展,2008,6(6):562-578.

[2] CARNEY K. Personalized medicine[J]. Journal of the California Dental Association,2003,4(6):548-558.

[3] GRIFFITH L G, SWARTZ M A. Capturing complex 3D tissue physiology in vitro[J]. Nature Reviews Molecular Cell Biology,2006,7(3):211-224.

[4] RICHMOND A, SU Y. Mouse xenograft models vs GEM models for human cancer therapeutics[J]. Disease Models & Mechanisms,2008,1(2/3):78-82.

[5] GARNETT M J, EDELMAN E J, HEIDORN S J, et al. Systematic identification of genomic markers of drug sensitivity in cancer cells[J]. Nature,2012,483(7391):570-575.

[6] SHOEMAKER R H. The NCI60 human tumour cell line anticancer drug screen[J]. Nature Reviews Cancer,2006,6(10):813-823.

[7] HEISER L M, WANG N J, TALCOTT C L, et al. Integrated analysis of breast cancer cell lines reveals unique signaling pathways[J]. Genome Biology,2009,10(3):31.

[8] KUTALIK Z, BECKMANN J S, BERGMANN S. A modular approach for integrative analysis of large-scale gene-expression and drug-response data[J]. Nature Biotechnology,2008,26(5):531-539.

[9] KANDELA I, ZERVANTONAKIS I. Registered report: Discovery and preclinical validation of drug indications using compendia of public gene expression data[J]. Science Translational Medicine,2011,3(96):9677-9687.

[10] ZHANG N, WANG H, FANG Y, et al. Predicting anticancer drug responses using a dual-layer integrated cell line-drug network model[J]. Plos Computational Biology,2015,11(9):1-4.

[11] KESHAVAN R H, OH S, MONTANARI A. Matrix completion from a few entries[J]. IEEE Transactions on Information Theory,2009,56(6):2980- 2998.

[12] CAI Z, ZHANG T, WAN X F. A computational framework for influenza antigenic cartography[J]. Plos Computational Biology,2010,6(10):1922-1928.

[13] BARRETINA J, CAPONIGRO G, STRANSKY N, et al. The Cancer Cell Line Encyclopedia enables predictive modeling of anticancer drug sensitivity[J]. Nature,2012,483(7391):603-607.

[14] VENKATESAN K, STRANSKY N, MARGOLIN A, et al. Prediction of drug response using genomic signatures from the Cancer Cell Line Encyclopedia[J]. Clinical Cancer Research,2010,16:2-5.

[15] DONG Z, ZHANG N, LI C, et al. Anticancer drug sensitivity prediction in cell lines from baseline gene expression through recursive feature selection[J]. BMC Cancer,2015,15(1):1-12.

[16] 申卯兴,郑武团.矩阵的奇异值分解的应用[J].大学数学,1996(3):56-61.

[17] 迟彬,叶庆凯.用奇异值分解方法计算具有重特征值矩阵的特征矢量[J].应用数学和力学,2004,25(3):233-238.

[18] PETERSEN K, PEDERSEN M. The Matrix Cookbook [M]. Copenhagen: Technical University of Denmark, 2012:12-14.

[19] BARNSTON A G. Correspondence among the correlation, RMSE, and Heidke forecast verification measures; refinement of the Heidke score[J]. Weather & Forecasting,2006,7(4):699-709.

[20] 牛晓太.基于KNN算法和10折交叉验证法的支持向量选取算法[J].华中师范大学学报(自科版),2014,48(3):335-338.

Matrix Completion for Prediction of the Cancerous Drug Sensitivity

HUANG Li, *HE Pingan*

(School of Sciences, Zhejiang Sci-Tech University, Hangzhou 310018, China)

Abstract: It is an important issue that how to use the incomplete data to obtain the complete and effective data, and then make the reliable analysis of the original data. In the work, based on a low-rank matrix completion, a novel low-rank matrix completion algorithm, which integrated the similarity information of gene expression data in cancer cell lines, was proposed to obtain reliable and complete anticancer drug sensitivity datasets. The model was applied to the observed datasets of cancer cell lines' responses to chemical compounds, recovering the missing data and re-evaluating the low react value. Compared to previous methods, the root-mean-square error (RMSE) in the method is reduced by 22.7% in a 10-fold cross validation analysis, which demonstrated that the novel algorithm can improve matrix completion quality.

Key words: anticancer drug sensitivity; low-rank matrix completion; cancer cell lines; 10-fold cross validation analysis; root-mean-square error (RMSE)

(责任编辑: 廖乾生)