基于最大后验概率估计的压缩感知算法

庄燕滨[1,2] 王尊志[1] 肖贤建[2] 张学武[3]
(河海大学计算机与信息学院 南京 211100)[1] (常州工学院计算机信息工程学院 常州 213002)[2]
(河海大学物联网工程学院 常州 213022)[3]

摘 要 针对压缩感知重构算法计算代价较大的问题,提出了一种用来构建压缩感知稀疏数据重构算法的MAP方法。此方法相对于一般的观测矩阵来说,计算代价较低。ℓ_1-范数使用一个标准的线性规划算法的最小计算代价是$O(N^3)$,该方法通过使用最大后验方法使计算代价减少到$O(N^2)$,并通过引入分割比来使算法更好地收敛。实验证明此方法能够获得较为成功的重构区域。

关键词 压缩感知,代价,后最大化,重构区域
中图法分类号 TP212 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2015.11.057

Reconstruction Algorithm in Compressed Sensing Based on Maximum Posteriori Estimation

ZHUANG Yan-bin[1,2] WANG Zun-zhi[1] XIAO Xian-jian[2] ZHANG Xue-wu[3]
(College of Computer & Information Engineering, Hohai University, Nanjing 211100, China)[1]
(School of Computer Information & Engineering, Changzhou Institute of Technology, Changzhou 213002, China)[2]
(College of Internet of Things Engineering, Hohai University, Changzhou 213022, China)[3]

Abstract This paper proposed a systematic method for constructing a sparse data reconstruction algorithm in compressed sensing which owns a relatively low computational cost for general observation matrix. It is known that the cost of ℓ_1-norm minimization using a standard linear programming algorithm is $O(N^3)$. The cost of proposed method can be reduced to $O(N^2)$ by applying the approach of posterior maximization. By introducing the division ratio, the algorithm achieves better convergence. Furthermore, in principle, the algorithm from our approach is expected to achieve the widest successful reconstruction region, which is evaluated from theoretical argument.

Keywords Compressed sensing, Cost, Posterior maximization, Reconstruction region

1 介绍

如今,压缩感知[1-3]方法已经被广泛应用于信息技术的各个领域,原始数据的稀疏性起了重要的作用[4],本文对压缩感知最基本的计算代价问题进行研究,首先考虑一个线性观测过程:

$$y=Fx \tag{1}$$

其中,$y\in R^M$是观测值,$x\in R^N$是含有许多零项的原始稀疏数据,其中非零项的个数由K表示,$M\times N$阶矩阵F描述了观测过程,并且M、N的大小被限制为一个压缩比率$\alpha=M/N<1$。再考虑最基本的ℓ_1-范数最小化问题:

$$\min_x \|x\|_1 \text{ 服从 } y=Fx \tag{2}$$

这是个线性规划问题,可以用一个计算代价为$O(N^3)$的标准算法求解,如内点法。此外,许多计算代价较小的替代算法[5,6]被提出,其代表为AMP算法。AMP(Approximate Message Passing)[7]算法是一个基于信息传递的阈值算法[8]。此算法中,矩阵F被假定为高斯随机矩阵[9],最终它的原始数据能被计算代价为$O(N^2)$的算法重构。此外,对于AMP算法,重构后的阈值在压缩率α和稀疏性$\beta(\beta=K/N)$方面与基于ℓ_1-范数[10]最小化问题的状态演变技术[7]得到的结果是相同的。然而,文献[7]中并没有解释AMP的阈值要等效于ℓ_1-范数最小化问题的原因。

本文提出一种用于构建稀疏数据重建算法的方法,此方法是对一个含有尽量小的M值的一般矩阵F使用最大后验概率(MAP)。通过此方法,可得到一个计算代价为$O(N^2)$的算法(特殊情况下,当矩阵F是稀疏的,并且在每一行/列只有O(1)个非零元素时,计算代价将减小到$O(N)$)。此外,本文的方法从另外一个角度解释了为什么AMP重构阈值和ℓ_1-范数最小化问题是等效的。

2 MAP重构算法

为了构建MAP算法,首先准备后验概率[11],从而从观测数据y和矩阵F推断原始数据x。定义后验概率为$P(x|y,F)$,其中x服从式(2)的ℓ_1-范数最小化问题。

$$P(x|y,F)=\lim_{\beta\to\infty}\frac{1}{Z(\beta)}\exp(-\beta\sum_i|x_i|)\prod_\mu\sigma(y_\mu-\sum_i F_{\mu i}x_i) \tag{3}$$

到稿日期:2014-11-02 返修日期:2015-01-24 本文受国家自然科学基金(61273170),江苏省科技厅工业支撑计划(BE2010072)资助。
庄燕滨 教授,硕士生导师,主要研究领域为智能信息处理、视频图像处理、模式识别;王尊志 硕士生,主要研究方向为图像处理;肖贤建 博士,副教授,主要研究方向为计算机视觉、模式识别;张学武 博士,副教授,主要研究方向为传感网与感知技术。

式中，$Z(\beta)=\exp(-\beta C)$，$C=\min_x \| x \|_1$，$1<i<N$，$1<\mu<M$。式(3)必须服从约束条件 $y=Fx$，否则概率为零。此外，只有当 ℓ_1-范数中 x 是最小值时它才满足，否则也为零。可知 MAP 的解决方案和 ℓ_1-范数最小值问题相似。注意，不需要限制 $\beta\to\infty$ 为 MAP 的解决方案，本文会用它来解释 MAP 算法和阈值重构分析之间的关系，通过使用统计力学的复制[12]方法，给出了阈值的精确描述[13]，文献中，后验概率被定义为波尔兹曼权重，极限 $\beta\to\infty$ 被作为分析的技术原因。

处理后验概率的一个难题是广义三角函数，将其调整为二次项形式：

$$P(x|y,F)=\lim_{\beta\to\infty}\frac{1}{Z(\beta)}\exp\left(-\beta\frac{\sum_\mu(y_\mu-\sum_i F_{\mu i}x_i)^2}{2}-\beta k\sum_i|x_i|\right) \tag{4}$$

式(4)通过约束 $\beta\to\infty$ 再现了 ℓ_1-范数最小值问题。在式(4)中，需要引入一个常量 $k(k>0)$ 来表示“相对意义”上的约束和最小化。要解决 ℓ_1-范数最小值问题，必须考虑极限约束条件 $k\to 0$，这个框架和文献[14]是基本相同的，需要对 MAP 算法采取适当的 $k\to 0$ 约束条件。

对于 MAP 方法，在 x_i 指数方面区分它们，稳态条件如下：

$$\sum_\mu F_{\mu i}y_\mu-\sum_{j(j\neq i)}\sum_\mu F_{\mu i}F_{\mu j}x_j-\sum_\mu F_{\mu i}^2 x_i-k\odot(x_i)=0 \tag{5}$$

其中，$\odot(x)$ 是 Heaviside[15] 函数，表示如下：

$$\begin{aligned}x_i&=\frac{1}{\sum_\mu F_{\mu i}^2}\eta\left(\sum_\mu F_{\mu i}(z_\mu+F_{\mu i}x_i);k\right)\\ z_\mu&=y_\mu-\sum_i F_{\mu i}x_i\end{aligned} \tag{6}$$

定义函数的阈值为：

$$\eta(x;k)=\begin{cases}x-k, & k<x\\ 0, & -k\leqslant x\leqslant k\\ x+k, & x<-k\end{cases} \tag{7}$$

变量 z_μ 表示 $y=Fx$ 约束的残余误差，可以通过求解稳态条件得到 ℓ_1-范数最小值。通过增加迭代步长上标 t 来构造一个迭代算法：

$$\begin{aligned}x_i^{(t)}&=\frac{1}{\sum_\mu F_{\mu i}^2}\eta\left(\sum_\mu F_{\mu i}(z_\mu^{(t)}+F_{\mu i}x_i^{(t-1)});k\right)\\ z_\mu^{(t)}&=y_\mu-\sum_i F_{\mu i}x_i^{(t-1)}\end{aligned} \tag{8}$$

MAP 的解决方案是找到固定点 $x_i^{(t)}$。ℓ_1-范数最小值问题是一个凸优化问题[16]，没有最小值固定点的位置。

下面给出一些关于式(8)算法的特征。

1)计算代价

在算法式(8)中只有单一的总和，$x_i^{(t)}$ 和 $z_\mu^{(t)}$ 的个数分别为 N 和 M。然而，当压缩率 M/N 是 O(1)时，计算代价为 $O(N^2)$，并且若矩阵 F 是稀疏的(每一行/列只有 O(1)个非零项，计算代价减小到 $O(N)$)，则收敛性的迭代步数比 N 和 M 更小。在一般情况下，它不在重建阈值的范围内，在收敛性和迭代步数方面 $k\to 0$ 收敛的速度必须放慢，因为 $k\to 0$ 在 N 和 M 值中占据主导地位。

2)矩阵 F 的适用性

本文不对 F 做任何假设，因此，在原则上该算法可以适用于一般的矩阵 F，而在原始 AMP 算法或者在统计力学分析[17]中矩阵 F 是假设值。

3)$k\to 0$ 极限

最终迭代步骤必须是极限 $k\to 0$，k 应该以 $k^{(t)}\propto\exp(-t/$常量)的方式指数衰减。这个衰减常量非常重要，如果常量太小会出现错误。这个常量在下文讨论的 AMP 算法中是一个阈值参数，对收敛性同样重要。在 AMP 算法最初的研究中，我们通过选择 k 均值平方误差对极限 $k\to 0$ 提出了适当的约束，并且与状态演化技术相结合，达到了收敛性的阈值重建，这在含有辅助变量 z 的 α 和 ρ 之间的关系中体现出来：

$$\rho=\alpha\max_{z\geqslant 0}\left(1-\frac{2}{\alpha}\{(1+z^2)H(z)-z\frac{e^{-z^2/2}}{\sqrt{2\pi}}\}/1+z^2-2\{(1+z^2)H(z)-z\frac{e^{-z^2/2}}{\sqrt{2\pi}}\}\right) \tag{9}$$

其中，$H(z)=\int_z^{+\infty}dx e^{-x^2/2}/\sqrt{2\pi}$ 相当于 ℓ_1-范数最小值，此公式根据文献[10,18]得出。下面同样给出相似方法的阈值方程：

$$\begin{aligned}&2(1-\rho)(H(z)-\frac{1}{z}\frac{e^{-z^2/2}}{\sqrt{2\pi}})+\rho=0\\ &\alpha=2(1-\rho)H(z)+\rho\end{aligned} \tag{10}$$

它们的解析也是等价的(可以通过代入数值检查式(9)和式(10)的等价性)。

算法的初始步骤是式(4)的概率分布，它是在阈值的分析中作为波尔兹曼权重。在 AMP 算法中，收敛条件式(10)已经用不含有 Onsager 条件[19]的阈值算法讨论过了，也是源于式(4)的分布。这种关系可以解释为什么有相同的阈值方程。

4)局部更新

阈值算法式(8)仍然存在问题，即不能总是通过式(8)得到一个正确的 ℓ_1-范数最小值，甚至不能从式(9)和式(10)得到重构区域，尽管极限 $k\to 0$ 设计得非常合适。这个错误发生在 ℓ_1-范数最小化重建阈值的地方。AMP 算法通过介绍 Onsager 的反应条件成功解决了这个问题，它是在原有的工作上减少微扰分析。本文从另外一个角度看待这个问题，并且适用于一般矩阵 F。

该策略如下，在原来的算法中，每步都更新变量 $x_i^{(t)}$。引入一个分割比，并修改算法来进行局部更新：

$$\begin{aligned}x_i^{(t)}&=\frac{\gamma^{(t)}}{1+\gamma^{(t)}}x_i^{(t)}+\frac{1}{1+\gamma^{(t)}}\frac{1}{\sum_\mu F_{\mu i}^2}\eta\left(\sum_\mu F_{\mu i}(z_\mu^{(t)}+F_{\mu i}x_i^{(t-1)});k\right)\\ z_\mu^{(t)}&=y_\mu-\sum_i F_{\mu i}x_i^{(t-1)}\end{aligned} \tag{11}$$

分割比 $\gamma^{(t)}$ 依赖于一般的步长，显然，这个算法的固定点和原来的相同；同时，通过适当地选择 $\gamma^{(t)}$ 改进了方法的收敛性。

在这里，$\gamma^{(t)}$ 可以任意选择，可以寻求如何设计 $\gamma^{(t)}$ 来使该算法的固定点的稳定分析有更好的收敛性。然而，目前还没有能选择 $\gamma^{(t)}$ 的较好的方法。

3 MAP 算法研究

以下是本文提出的算法的特性。矩阵 F[20]是从含有零均值并且方差为 $1/M$ 的高斯分布上得出的独立同分布项，它在 AMP 算法和统计力学分析[18]中是一样的。对于 M、N，有 $\lim_{M\to\infty}\sum_\mu F_{\mu i}^2=1$，式(8)简化为：

$$x_i^{(t)}=\eta(\sum_\mu F_{\mu i}z_\mu^{(t)}+x_i^{(t-1)};k)$$
$$z_\mu^{(t)}=y_\mu-\sum_i F_{\mu i}x_i^{(t-1)} \tag{12}$$

式(12)在文献[7]中也有介绍，其中提到式(8)的阈值算法的效果不如 ℓ_1-范数最小值的效果好。在式(12)第一个公式中引入一个分割比来进行改善(这里 γ 不依赖于 t)：

$$x_i^{(t)}=\frac{\gamma}{1+\gamma}x_i^{(t-1)}+\frac{1}{1+\gamma}\eta(\sum_\mu F_{\mu i}z_\mu^{(t)}+x_i^{(t-1)};k) \tag{13}$$

根据 η 的特性上式可以写成(重新调整参数 k，它和固定点无关，因为最终 k 会变成 0)：

$$x_i^{(t)}=\frac{\gamma}{1+\gamma}x_i^{(t-1)}+\eta(\frac{1}{1+\gamma}\{\sum_\mu F_{\mu i}z_\mu^{(t)}+x_i^{(t-1)}\};k) \tag{14}$$

在式(14)上修改一下更新规则，得：

$$x_i^{(t)}=\eta(\frac{1}{1+\gamma}\sum_\mu F_{\mu i}z_\mu^{(t)}+x_i^{(t-1)};k) \tag{15}$$

等号右边通过变换最终与 η 相关。这与式(14)的固定点一样(可以通过消除上标 t 验证)。重新调整残余误差 $\hat{z}_\mu^{(t)}=z_\mu^{(t)}/(1+\gamma)$ 可以得到：

$$x_i^{(t)}=\eta(\sum_\mu F_{\mu i}\hat{z}_\mu^{(t)}+x_i^{(t-1)};k)$$
$$\hat{z}_\mu^{(t)}=\frac{1}{1+\gamma}\{y_\mu-\sum_i F_{\mu i}x_i^{(t-1)}\} \tag{16}$$

式(16)表明，分配比例与残差 z_μ 的尺度相关。接下来，考虑步依赖 $\gamma^{(t)}$。

$$\gamma^{(t)}=\frac{1}{M}\sum_i\eta'(\sum_\mu F_{\mu i}z_\mu^{(t)}+x_i^{(t-1)};k) \tag{17}$$

其中 $\eta'(x;k)=\partial_x\eta(x;k)$，$\gamma^{(t)}$ 的选择是为了在求解二阶稳定分析中实现更快的收敛速度。$\gamma^{(t)}$ 是由阈值函数表示，即式(7)。当重构算法成功时，$\gamma^{(t)}$ 的值是 $K/M=\rho/\alpha(t\to\infty)$。对于步依赖 $\gamma^{(t)}$，通过式(7)也可以得到算法的另外一种表示形式：

$$x_i^{(t)}=\eta(\sum_\mu F_{\mu i}\hat{z}_\mu^{(t)}+x_i^{(t-1)};k)$$
$$z_\mu^{(t)}=y_\mu-\sum_i F_{\mu i}x_i^{(t-1)}$$
$$\hat{z}_\mu^{(t)}=\frac{1}{1+\frac{1}{M}\sum_j\eta'(\sum_\mu F_{\mu j}z_\mu^{(t)}+x_j^{(t-1)};k)}z_\mu^{(t)} \tag{18}$$

式(18)表明，残差 z_μ 的尺度随着 t 而变化。

4 与 AMP 算法的联系

根据上面的观测，取分割比 $\gamma^{(t)}$ 为负值：

$$\gamma^{(t)}=-\frac{1}{M}\sum_i\eta'(\sum_\mu F_{\mu i}z_\mu^{(t)}+x_i^{(t-1)};k) \tag{19}$$

这会导致局部更新。这种状况下，有与式(18)相似的公式，唯一的不同是第 3 个等式里面分母部分是负值。这里，介绍分割比 $\gamma^{(t)}$ 在更新残差方面的调整：

$$\hat{z}_\mu^{(t)}=\hat{\gamma}^{(t-1)}\hat{z}_\mu^{(t-1)}+\hat{z}_\mu^{(t)}(1-\hat{\gamma}^{(t-1)})/(1-\frac{1}{M}\sum_j\eta'$$
$$(F_{\mu j}z_\mu^{(t)}+x_j^{(t-1)};k)) \tag{20}$$

选择 $\hat{\gamma}^{(t)}=-\gamma^{(t)}=\sum_j\eta'(\sum_\mu F_{\mu j}z_\mu^{(t)}+x_j^{(t-1)})/M$，这个算法可以写成：

$$x_i^{(t)}=\eta(\sum_\mu F_{\mu i}\hat{z}_\mu^{(t)}+x_i^{(t-1)};k)$$
$$z_\mu=y_\mu-\sum_i F_{\mu i}x_i$$
$$\hat{z}_\mu^{(t)}=z_\mu^{(t)}+\frac{1}{M}\hat{z}_\mu^{(t-1)}\sum_j\eta'(\sum_\mu F_{\mu j}z_\mu^{(t)}+x_j^{(t-1)};k) \tag{21}$$

通过引入分割比和重写的更新规则得到这个表述，当把上式中的第 3 个等式里的 $z_\mu^{(t)}$ 替代为 $\hat{z}_\mu^{(t-1)}$(近似地 $z\to\hat{z}$，固定点不变，因为它们最后都会达到零)，并把 η' 中 x 项的上标 $t-1$ 变成 $t-2$，最后得到

$$x_i^{(t)}=\eta(\sum_\mu F_{\mu i}\hat{z}_\mu^{(t)}+x_i^{(t-1)};k)$$
$$\hat{z}_\mu^{(t)}=y_\mu-\sum_i F_{\mu i}x_i^{(t-1)}+\frac{1}{M}\hat{z}_\mu^{(t-1)}\sum_j\eta'(\sum_\mu F_{\mu j}\hat{z}_\mu^{(t-1)}+$$
$$x_j^{(t-2)};k) \tag{22}$$

这只是 AMP 算法中变量的更新，其中参数 k 控制阈值。因此，可以发现 AMP 和部分更新 MAP 算法之间的联系。在 AMP 算法中，式(22)中第 2 个等式中的最后一项是来自消息传递的探讨，并在统计力学[7,13]中被解释为 Onsager 反应。根据这个逻辑可得到一个结论，通过一个分割比能实现算法更好的收敛性。

5 数值实验

5.1 重构阈值

首先，评估本文算法的重构阈值。在实验中，原数据维数 $N=10^3$ 是固定的，观测值 M 和非零数据集 K 是变化的。F 的每一项及其零均值和方差 $1/M$ 在高斯分布中是随机和独立的(见图 1—图 3)，每个非零项及其零均值和方差同样服从高斯分布。

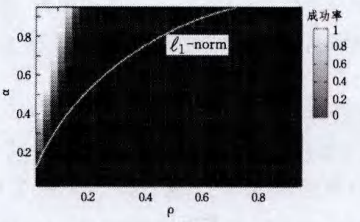

图 1　不含分割比的 MAP 算法和 ℓ_1-范数结果比较

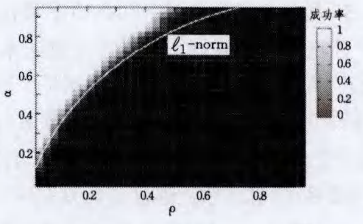

图 2　分割比为 1 的 MAP 算法和 ℓ_1-范数结果比较

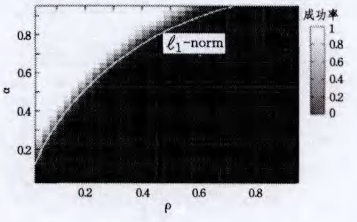

图 3　分割比为式(17)的 MAP 算法和 ℓ_1-范数结果比较

初始值 $x_i^{(t)}$ 设置为零，在每一步更新中 k 以指数乘以 0.999 衰减，所以 k 会慢慢变为零，但是收敛速度很慢。在图 1 和图 2 中，是 5×10^3 步之后的 $x_i^{(t)}$ 重构的结果，在图 3 和图 4 中是 10^4 步。为了评价重构阈值，进行了 50 次实验，并在

M、K 固定时计算成功率。在每次实验中，重构是否成功的判断准则是每次数据的均方误差是否小于 10^{-3}。然后，通过改变每一对 M、K 的值来计算成功率(每次实验 M、K 的值增加或者减少 25)。

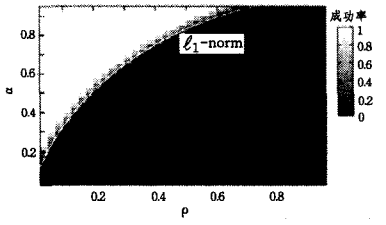

图 4　随机矩阵 F 产生的 MAP 算法和 ℓ_1-范数结果比较

实验结果见图 1—图 3，实验中获得的成功率在灰度图中表示出来。在图 1 中，算法的结果不含有分割比，较 ℓ_1-范数最小化的成功重构区域窄。图 2 是分割比 $\gamma^{(t)}=1$ 的结果，重构阈值接近 ℓ_1-范数阈值曲线。为了提高效果，采用式(17)中的分割比，这一修改可以获得更接近 ℓ_1-范数的曲线，如图 3 所示，但在较低的压缩比区域，效果没有 ℓ_1-范数好。然而，更多的迭代步数、适当地选择 k 的值和 $\gamma^{(t)}$ 会取得更好的效果。下面为了检测一般矩阵 F 的适用性，使随机矩阵 F 由不同方式产生：首先，产生一个随机矩阵 F，然后随机消除 90% 的项，这里设置为零。在这个实验中使用式(11)，分割比 $\gamma^{(t)}$ 为 1，图 4 的结果显示，成功区域和 ℓ_1-范数相似。

从这个结果确认，MAP 算法本质上等价于 ℓ_1-范数，并且，通过一个合适的算法设计，可以得到和 ℓ_1-范数几乎相同的效果，需要强调的是，这里的算法也在一般矩阵 F 下适用，如图 4 所示。

5.2　算法的收敛性

本节研究算法收敛速度，并比较两个算法：分割比为式(17)的 MAP 算法和 AMP 算法。设 $N=2000$，$M=1000$(压缩率 $\alpha=0.5$)，$K=200$(非零项比例 $\rho=0.1$)。这个方案产生的矩阵 F 和初始数据 x^0 与图 1—图 3 是相同的。参数 k 以每步乘以 0.95 指数衰减到零(最初的 AMP 算法是以均方误差更新 k，并且收敛性比较好。这里两个算法采用缓慢的指数更新方法)。进行 100 次重构实验，观测每个数据的均方误差结果，如图(5)所示。能够看出与 MAP 算法相比，AMP 算法的均方误差较小，且收敛速度较快，从上文讨论的 MAP 和 AMP 算法的联系中，算法收敛速度的区别在于 AMP 算法考虑了残差 z_μ 的分割比，而 MAP 算法(11)没有考虑，因此，有必要考虑更快的残差 z_μ 的收敛速度来改进 MAP 算法。

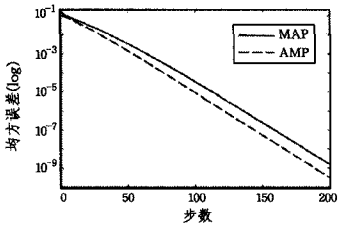

图 5　AMP 算法和 MAP 算法收敛速度比较

结束语　本文提出了 MAP 重构算法，讨论了其与已知阈值算法的联系，并且通过数值实验评估了它的性能。证明了通过设计适当的算法可以取得与 ℓ_1-范数最小化几乎相同的重构阈值。最重要的是，该算法不对矩阵 F 做任何假设，因此算法重构计算代价相对较低，并且适用于一般矩阵。

本文同时提出要找到更合适的分割比使 MAP 算法收敛速度更快。因此未来的工作是采用本文的讨论设计一个更好的方法来实现更快的收敛速度。

参 考 文 献

[1] Donoho D L. Compressed sensing[J]. IEEE Transactions on Information Theory, 2006, 52(4): 1289-1306

[2] Candes E J, Tao T. Decoding by linear programming[J]. IEEE Transactions on Information Theory, 2005, 51(12): 4203-4215

[3] Candes E J, Romberg J, Tao T. Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information[J]. IEEE Transactions on Information Theory, 2006, 52(2): 489-509

[4] Candes E J, Wakin M B. An introduction to compressive sampling[J]. Signal Processing Magazine, IEEE, 2008, 25(2): 21-30

[5] Elad M, Figueiredo M A T, Ma Yi. On the role of sparse and redundant representations in image processing[J]. Proceedings of the IEEE, 2010, 98(6): 972-982

[6] Tropp J A, Wright S J. Computational methods for sparse solution of linear inverse problems[J]. Proceedings of the IEEE, 2010, 98(6): 948-958

[7] Donoho D L, Maleki A, Montanari A. Message passing algorithms for compressed sensing: i. motivation and construction [C]// Information Theory Workshop (ITW). 2010 IEEE, 2010: 1-5

[8] Kj S, Andersen R. Probabilistic reasoning in intelligent systems: networks of plausible inference judea pearl[J]. Artificial Intelligence, 1991, 4(1): 117-124

[9] 赵玉娟，郑宝玉，陈守宁. 压缩感知自适应观测矩阵设计[J]. 信号处理，2012，28(12)：1635-1641
Zhao Yu-juan, Zheng Bao-yu, Chen Shou-ning. The design of adaptive measurement matrix in compressed sensing[J]. Signal Processing, 2012, 28(12): 1635-1641

[10] Recht B, Xu Wei-yu, Hassibi B. Null space conditions and thresholds for rank minimization[J]. Mathematical Programming, 2011, 127(1): 175-202

[11] 冯祖仁，吕娜，李良福. 基于最大后验概率的图像匹配相似性指标研究[J]. 自动化学报，2007，33(1)：1-8
Feng Zhu-ren, Lv Na, Li Liang-fu. Research on Image Matching Similarity Criterion Based on Maximun Posterior Probability [J]. Acta Automation Sinica, 2007, 33(1): 1-8

[12] Tibshirani R. regression shrinkage and selection via the lasso [J]. Journal of the Royal Statistical Society, series b: methodological, 1996, 58(1): 267-288

[13] Bayati M, Montanari A. The dynamics of message passing on dense graphs, with applications to compressed sensing[J]. IEEE Transactions on Information Theory, 2011, 57(2): 764-785

[14] Krzakala F, Mézard M, Sausset F. Probabilistic reconstruction in compressed sensing: algorithms, phase diagrams, and threshold achieving matrices[J]. Journal of Statistical Mechanics: Theory and Experiment, 2012(4): P08009

[15] 李书波，张渡淮. heaviside 函数的非标准分析表示[J]. 哈尔滨科学技术大学学报，1987，45(4)：760-782
Li Shu-bo, Zhang Du-huai. Non-standard analysis representation

of heaviside function[J]. Journal of Harbin University of Science and Technology,1987,45(4):760-782

[16] 李慧玲,张春阳,李卓识,等.解非凸优化问题的一个同伦内点方法[J].东北师大学报(自然科学版),2009,41(4):764-785
Li Hui-ling,Zhang Chun-yang,Li Zuo-shi,et al. Homotopy interior point method for non-convex optimization problem with weak pseudo cone condition[J]. Journal of Northeast Normal University(Natural Science Edition),2009,41(4):764-785

[17] Gupta S. A note on the asymptotic distribution of lasso estimator for correlated data[J]. Sankhya A,2012,74(1):10-28

[18] Kutyniok G. Theory and applications of compressed sensing[J]. Gamm-mitteilungen,2013,36(1):79-101

[19] Gallavotti G. Chaotic hypothesis:onsager reciprocity and fluctuation-dissipation theorem[J]. Journal of Statistical Physics,1996,84(5/6):899-925

[20] 王定成,苏淳. b 值独立同分布随机变元序列的矩完全收敛性[J].应用数学学报,2004,27(3):440-448
Wang Ding-cheng, Sun Chun. Moment Complete Convergence for Sums of i. i. d. Random Elements in Banach Space[J]. Acta Mathematicae Applicatae Sinica,2004,27(3):440-448

(上接第 255 页)

可以发现 GKNNimpute 算法耗时最多,RKNNimpute 和 IKNNimpute 算法相对快捷。

综上,RKNNimpute 与 GKNNimpute 相比,二者有近似相等的填补准确率,RKNNimpute 具有更小的计算复杂度。通过比较基于精简关联度迭代算法(RKNNimpute)和基于灰色关联度迭代算法(GKNNimpute)的填补性能(两算法仅相似性的度量方法不同,其他细节均相同),间接比较精简关联度(RRG)和灰色关联度(GRG)的性能。在相似性衡量上,精简关联度能达到与灰色关联度完全相同的效果;在时间开销上,精简关联度大大降低了算法的时间复杂度。因而,本文提出的精简关联度是一种优秀的距离度量方法。与 IKNNimpute 算法比较可知,RKNNimpute 算法准确率更高,针对基因表达数据能取得不错的填补效果。

结束语 本文提出了新的关联度计算方案——精简关联度,它是对灰色关联度的改进。选用 3 种类型的数据集,设置不同的缺失率,进行大量实验,分析了精简关联度的性能。在预测准确性上,它与灰色关联度有一致的相似性度量效果;在时间复杂度上,所提出的度量方法显著地降低了时间复杂度。因而,本文提出的精简关联度是一种高效的相似性度量方法。在此基础上,针对基因表达数据的缺失现象,进一步提出了基于精简关联度的迭代填补算法。RKNNimpute 利用填补后的基因和完整基因构成候选基因集,扩大了近邻的选择范围,提高了数据的利用率;通过设置合适的阈值,以迭代的方式提高了填补性能。与迭代 K 近邻填补算法进行了比较,实验证实 RKNNimpute 是一种有效的填补算法。

RKNNimpute 算法需预先指定迭代终止的条件,即收敛的精度和最大迭代次数,它们能影响算法的迭代次数和填补准确率。如何高效合理地确定收敛精度和最大迭代次数以及进一步提高算法的性能是今后的研究方向。

参 考 文 献

[1] Hoheisel J D. Microarray technology:beyond transcript profiling and genotype analysis [J]. Nature Reviews Genetics, 2006, 7(3):200-210

[2] De Brevern A G, Hazout S, Malpertuy A. Influence of microarrays experiments missing values on the stability of gene groups by hierarchical clustering [J]. BMC Bioinformatics, 2004, 5(1): 114-119

[3] Yang Y H, Buckley M J, Dudoit S, et al. Comparison of methods for image analysis on cDNA microarray data [J]. Journal of Computational and Graphical Statistics, 2002, 11(1): 108-136

[4] Pedro J, Garcia-Laencina, et al. K nearest neighbours with mutual information for simultaneous classification and missing data imputation [J]. Neurocomputing, 2009, 72(7-9): 1483-1493

[5] Moorthy K, Mohamad M S, Deris S. A Review on Missing Value Imputation Algorithms for Microarray Gene Expression Data [J]. Current Bioinformatics, 2014, 9(1): 18-22

[6] Song Qin-bao, Shepperd M, Chen Xiang-ru, et al. Can k-NN imputation improve the performance of C4. 5 with small software project data sets? A comparative evaluation [J]. Journal of Systems and Software, 2008, 81(12): 2361-2370

[7] Troyanskaya O, Cantor M, Sherlock G. Missing value estimation methods for DNA microarrays [J]. Bioinformatics, 2001, 17(6): 520-525

[8] Alan Wee-Chung, Law Ngai-Fong, Yan Hong. Missing value imputation for gene expression data: computational technique to recover missing data from available information [J]. Briefings in Bioinformatics, 2010, 12(5): 498-513

[9] Meng Fan-chi, Cheng Cai, Hong Yan. A Bicluster-Based Bayesian Principal Component Analysis Method for Microarray Missing Value Estimation [J]. Biomedical and Health Informatics, 2014, 18(3): 862-871

[10] Zhang Shi-chao. Shell-neighbor method and its application in missing data [J]. Applied Intelligence, 2011, 35(1): 123-133

[11] 杨涛.基因表达缺失数据填充算法研究[D].长沙:湖南大学,2005
Yang Tao. The research on imputation algorithm of missing values for gene expression data [D]. Changsha: Hunan University, 2005

[12] Zhang Shi-chao. NIIA: Nonparametric Iterative Imputation Algorithm[C] // Trends in Artificial Intelligence, 2008 (PRICAI 2008). Berlin: Springer Berlin Heidelberg, 2008: 544-555

[13] Song Qin-bao, Shepperd M. Predicting software project effort: A grey relational analysis based method [J]. Expert Systems with Applications, 2011, 38(6): 7302-7316

[14] Bras L P, Menezes J C. Improving cluster-based missing value estimation of DNA microarray data [J]. Biomolecular Engineering, 2007, 24(2): 273-282

[15] 李艳芳.基因表达数据的缺失值估计研究[D].哈尔滨:哈尔滨工业大学,2011
Li Yan-fang. Research on missing value imputation for microarray gene expression data [D]. Harbin: Harbin Institute of Technology, 2011