两两关系马尔科夫网的自适应组稀疏化学习

刘建伟 [1] 任正平 [1] 刘泽宇 [2] 黎海恩 [1] 罗雄麟 [1]

摘 要 稀疏化学习能显著降低无向图模型的参数学习与结构学习的复杂性, 有效地处理无向图模型的学习问题. 两两关系马尔科夫网在多值变量情况下, 每条边具有多个参数, 本文对此给出边参数向量的组稀疏化学习, 提出自适应组稀疏化, 根据参数向量的模大小自适应调整惩罚程度. 本文不仅对比了不同边势情况下的稀疏化学习性能, 为了加速模型在复杂网络中的训练过程, 还对目标函数进行伪似然近似、平均场自由能近似和 Bethe 自由能近似. 本文还给出自适应组稀疏化目标函数分别使用谱投影梯度算法和投影拟牛顿算法时的最优解, 并对比了两种优化算法进行稀疏化学习的性能. 实验表明自适应组稀疏化具有良好的性能.

关键词 无向图模型, 两两马尔科夫网, 稀疏化学习, 自适应组稀疏化

引用格式 刘建伟, 任正平, 刘泽宇, 黎海恩, 罗雄麟. 两两关系马尔科夫网的自适应组稀疏化学习. 自动化学报, 2015, **41**(8): 1419−1437

DOI 10.16383/j.aas.2015.c140682

Adaptive Group Sparse Learning of Pairwise Markov Network

LIU Jian-Wei[1] REN Zheng-Ping[1] LIU Ze-Yu[2] LI Hai-En[1] LUO Xiong-Lin[1]

Abstract Sparse learning can significantly reduce the complexity of parameter learning and structure learning and effectively deal with learning problems of undirected graphical models. In the case of pairwise Markov network, in which each variable has more than two values, the number of parameters associated with an edge is more than one. This paper proposes a group sparse learning approach for the parameters associated with edges, and puts forward an adaptive group sparse learning algorithm, which can adaptively adjust the degree of penalty according to the norm of the parameters vector. This paper compares the performance of sparse learning using different edge potentials. In order to speed up the training process, three approximate object functions are given, including pseudo likelihood approximation, mean field approximation and Bethe free energy approximation. Two optimization algorithms, i. e., projected quasi-Newton algorithm and spectral projected gradient algorithm, are also compared. Experimental results show that the proposed adaptive group sparse learning algorithm outperforms the normal sparse learning ones.

Key words Undirected graphical models, pairwise Markov network, sparse learning, adaptive group sparsity

Citation Liu Jian-Wei, Ren Zheng-Ping, Liu Ze-Yu, Li Hai-En, Luo Xiong-Lin. Adaptive group sparse learning of pairwise Markov network. *Acta Automatica Sinica*, 2015, **41**(8): 1419−1437

稀疏化学习得到的稀疏模型具有更少的参数个数, 使得参数估计的计算成本更低、效率更高. 稀疏模型在学习过程中得到的边集能够直观地反映数据集中所描述的变量之间的关系, 得到网络结构. 文献 [1−3] 对概率图模型的表示、学习以及稀疏化学习进行了详细的阐述. 本文主要研究的是两两关系马尔科夫网在多值变量情况下的稀疏化学习问题.

收稿日期 2014-09-24 录用日期 2015-02-16
Manuscript received September 24, 2014; accepted February 16, 2015
中国石油大学 (北京) 基础学科研究基金项目 (JCXK-2011-07) 资助
Supported by Foundation Sciences China University of Petroleum (JCXK-2011-07)
本文责任编委 刘成林
Recommended by Associate Editor LIU Cheng-Lin
1. 中国石油大学自动化研究所 北京 102249 2. 中国科学院软件研究所基础软件国家工程研究中心 北京 100190
1. Research Institute of Automation, China University of Petroleum, Beijing 102249 2. National Engineering Research Center for Fundamental Software, Institute of Software, Chinese Academy of Sciences, Beijing 100190

在多值变量情况下, 两两关系马尔科夫网中某一条边的两个节点之间, 不同的取值组合会对边的参数产生不同影响. 此时, 一条边与多个参数相关, 稀疏学习时, 必须使与这条边相关的所有参数为 0, 才能从网络图中将该条边删除. 稀疏化学习的过程中需要对边的参数向量进行组共同稀疏化[4−6], 普通的 L_1 正则化稀疏模型不再适用于多值变量场景. 另外, 普通组稀疏对每条边的参数向量的惩罚不会随着向量的模的大小而改变, 会对模较大的参数向量过度缩小, 导致统计估计有偏性.

因此, 本文不仅给出了两两关系马尔科夫网的组稀疏化学习, 为了避免估计有偏性, 本文还提出两两关系马尔科夫网的自适应组稀疏模型, 为不同参数向量分配不同的权, 对模较大的参数向量执行程度较小的惩罚, 而对模较小的参数向量执行程度较大的惩罚.

本文的主要工作有:

1) 本文在两两关系马尔科夫网[7−11] 的概率密度模型, 加上一个对模型参数进行 L_1 正则化的罚项[12] 的基础上提出了自适应组稀疏化模型[13−14], 对边的参数向量进行组共同稀疏化, 并且避免了估计有偏性[15−16].

2) 本文在人工数据集和实际数据集上对比了不同边势情况下不同稀疏模型的性能, 实验表明完全边势情况下的自适应组稀疏化模型性能最好.

3) 为了使模型能够运用在复杂网络结构中, 本文对目标函数进行了三种近似: 伪似然近似[15−16]、平均场自由能近似[17] 和 Bethe 自由能近似[18−20]. 三种近似目标函数中, 性能最佳的是伪似然近似, 另外两种近似方法虽然运行时间大大缩短, 其对模型的拟合性能和稀疏性能也大大降低. 同样, 在近似目标函数的人工数据集和实际数据集实验中, 自适应组稀疏化学习仍具有更好的性能.

4) 本文给出了自适应组稀疏模型使用谱投影梯度算法和投影拟牛顿法的最优解, 并对比了两种优化算法的性能. 实验表明, 在实际数据集上, 谱投影梯度算法的拟合性能和运行时间优于投影拟牛顿法. 不论在人工数据集还是实际数据集上, 使用自适应组稀疏化模型性能更优.

本文具体安排如下: 第 1 节介绍两两关系马尔科夫网和不同边势形式, 给出自适应组稀疏化模型; 第 2 节对精确目标函数进行三种不同近似: 伪似然近似、平均场近似和 Bethe 自由能近似; 第 3 节给出了两种不同的优化算法: 谱投影梯度法和投影拟牛顿法, 并给出了这两种优化算法在自适应组稀疏化时的最优解; 第 4 节在人工数据集和实际数据集上, 对比不同正则化模型在使用不同边势、不同近似目标函数和不同优化算法时的性能; 最后, 第 5 节对全文进行总结和展望.

1 两两关系马尔科夫网的自适应组稀疏化

1.1 两两关系马尔科夫网与不同边势形式

假设有 p 个随机变量 $\boldsymbol{X} = (X_1, \cdots, X_p)$, $G = (V, E)$ 为 p 个随机变量的两两关系马尔科夫网, 其中 V 为节点集, E 为边集. 两两关系马尔科夫网只考虑单个节点和有边相连的两个节点的相互作用[7−11], 因此其联合概率分布只包含单节点势能 $\{\phi(X_i) : i = 1, \cdots, p\}$ 和边势能 $\{\phi(X_i, X_j) : (i, j) \in E\}$. 其联合概率分布的对数线性模型表示如下:

$$P(\boldsymbol{X}) = \frac{1}{Z(\boldsymbol{\omega}, \boldsymbol{b})} \prod_{i=1}^{p} \phi_i(X_i, b_i) \prod_{(i,j)\in E} \phi_{ij}(X_i, X_j, \omega_{ij}) \quad (1)$$

向量 $\boldsymbol{b}$ 表示所有节点参数的集合, 向量 $\boldsymbol{\omega}$ 表示所有边对应的参数的集合, $Z(\boldsymbol{\omega}, \boldsymbol{b}) = \sum_{\boldsymbol{X}} \prod_{i=1}^{p} \phi_i(X_i, b_i) \prod_{(i,j)\in E} \phi_{ij}(X_i, X_j, \omega_{ij})$ 为归一化函数. 本节讨论的模型满足对数线性, 因此把式 (1) 改写成负对数似然形式. 当给出随机变量集 $\boldsymbol{X}$ 的 n 个样本时, 其负对数似然函数为

$$-\sum_{m=1}^{n}\Bigg[\sum_{i=1}^{p}\Bigg[\log\phi_i(x_i^m, b_i) + \sum_{j=i+1}^{p}\log\phi_{ij}(x_i^m, x_j^m, \omega_{ij})\Bigg)\Bigg]\Bigg] + n\log Z(\boldsymbol{\omega}, \boldsymbol{b}) \quad (2)$$

由于在学习过程中事先并不知道网络结构, 因此上式的边势考虑了所有节点两两之间的相互作用, 此时的 ω_{ij} 扩展为所有节点两两之间的边权. 若 ω_{ij} 取值为 0, 则节点 X_i 和 X_j 之间不存在边连接.

在多值变量情况中, 假设变量集 $\boldsymbol{X}$ 的取值 $\boldsymbol{x} \in \{1, 2, \cdots, k\}^p$, 即每个变量 X_i 有 k 种取值状态. 此时, 每个势能函数的对数形式都能表示为关于参数的线性函数. 例如, 假设变量 X_i 有 4 种取值状态 ($k = 4$), 则单节点势能 $\phi_i(x_i)$ 的对数形式可写成:

$$\log\phi_i(x_i, \boldsymbol{b}_i) = \mathrm{I}(x_i = 1)\, b_{i,1} + \mathrm{I}(x_i = 2)\, b_{i,2} + \mathrm{I}(x_i = 3)\, b_{i,3} \quad (3)$$

其中, b_{ik} 表示与节点 X_i 的第 k 种取值状态有关的参数, $\mathrm{I}(\cdot)$ 为指示函数, 若其自变量为真, 则返回 1; 否则, 返回 0. 由于全局归一化的过程会对每个势能函数重新调整, 所以对于具有 k 种取值状态的节点, 只需使用 $k-1$ 个参数来表示其节点势能. 节点势能的参数集表示为 $\boldsymbol{b}_i = (b_{i,1}, b_{i,2}, \cdots, b_{i,k-1})$ 而边势能 $\phi_{ij}(x_i, x_j)$ 的边权 ω_{ij} 在多值变量的情况中则比较复杂. 在多个取值状态下, 认为两个节点的不同取值组合对边权产生不同程度的影响. 那么, 每一条边根据不同的取值组合情况而具有多个参数, 只有当该边的所有参数都为 0 时, 才能认为这一条边不存在. 两个节点的取值组合主要可分为三种情况.

1) 只有当两个节点都取相同值时才对边权产生影响, 而且无论两个节点取何值时对边权产生的影响都是一样的. 例如, 当 $k = 3$ 时的边势能的对数形式为

$$\log\phi_{ij}(x_i, x_j, \omega_{ij}) = \mathrm{I}(x_i = 1, x_j = 1)\,\omega_{ij} + \mathrm{I}(x_i = 2, x_j = 2)\,\omega_{ij} + \mathrm{I}(x_i = 3, x_j = 3)\,\omega_{ij} \quad (4)$$

把上式表示为矩阵形式

$$\log\phi_{ij}\left(x_i,x_j,\omega_{ij}\right)=I\left[\begin{array}{ccc}\omega_{ij}&0&0\\0&\omega_{ij}&0\\0&0&\omega_{ij}\end{array}\right]\tag{5}$$

其中, I 为单位矩阵. 类似于 Ising 模型的二值变量情况, 对数边势矩阵对角线上的元素全部相同. 只要取值为 0, 即可删去该边. 本节把这种边势称为对角相同边势.

2) 只有当两个节点都取相同值时才对边权产生影响, 而且两个节点的取值的不同会对边权产生不同的影响. 例如, 当 $k=3$ 时的边势能的对数形式为

$$\begin{aligned}&\log\phi_{ij}\left(x_i,x_j,\boldsymbol{\omega}_{ij}\right)=\mathrm{I}\left(x_i=1,x_j=1\right)\omega_{ij1}+\\&\quad\mathrm{I}\left(x_i=2,x_j=2\right)\omega_{ij2}+\mathrm{I}\left(x_i=3,x_j=3\right)\omega_{ij3}\end{aligned}\tag{6}$$

把上式表示为矩阵形:

$$\log\phi_{ij}\left(x_i,x_j,\boldsymbol{\omega}_{ij}\right)=I\left[\begin{array}{ccc}\omega_{ij1}&0&0\\0&\omega_{ij2}&0\\0&0&\omega_{ij3}\end{array}\right]\tag{7}$$

向量 $\boldsymbol{w}_{ij}$ 表示边势的参数集, 根据两个节点共同取值的不同而相应给出不同的边权, 使得对数边势矩阵对角线上的元素各不相同. 只有当向量 $\boldsymbol{\omega}_{ij}$ 中的所有参数都同时为 0, 才能认为节点 X_i 和 X_j 之间没有边连接. 本节把该边势称为对角不同边势.

3) 是最复杂的情况, 两个节点的所有取值组合都对边权产生不同的影响. 例如, 当 $k=3$ 时的边势能的对数形式为

$$\begin{aligned}&\log\phi_{ij}\left(x_i,x_j,\omega_{ij}\right)=\mathrm{I}\left(x_i=1,x_j=1\right)\omega_{ij11}+\\&\mathrm{I}\left(x_i=1,x_j=2\right)\omega_{ij12}+\mathrm{I}\left(x_i=1,x_j=3\right)\omega_{ij13}+\\&\mathrm{I}\left(x_i=2,x_j=1\right)\omega_{ij21}+\mathrm{I}\left(x_i=2,x_j=2\right)\omega_{ij22}+\\&\mathrm{I}\left(x_i=2,x_j=3\right)\omega_{ij23}+\mathrm{I}\left(x_i=3,x_j=1\right)\omega_{ij31}+\\&\mathrm{I}\left(x_i=3,x_j=2\right)\omega_{ij32}+\mathrm{I}\left(x_i=3,x_j=3\right)\omega_{ij33}\end{aligned}\tag{8}$$

把上式表示为矩阵形式

$$\log\phi_{ij}\left(x_i,x_j,\omega_{ij}\right)=I\left[\begin{array}{ccc}\omega_{ij11}&\omega_{ij12}&\omega_{ij13}\\\omega_{ij21}&\omega_{ij22}&\omega_{ij23}\\\omega_{ij31}&\omega_{ij32}&\omega_{ij33}\end{array}\right]\tag{9}$$

w_{ij} 表示边势的参数矩阵, 对两个节点的所有取值组合都给出不同的边权, 使得对数边势矩阵的每个元素都不同, 称这种边势为完全边势[17]. 因此, 当节点具有 k 种取值状态时, 每条边连接的两个不同节点均有 k 种取值, 那么该条边的边势都具有 k^2 个参数. 只有当这 k^2 个参数都为 0 时, 才认为该边不属于图模型的边集. 本节讨论的自适应组稀疏化学习主要针对完全边势.

1.2 两两关系马尔科夫网的自适应组稀疏化模型

对两两关系马尔科夫网的多值情况来说, 组就是与某一条边相关的所有参数构成的向量. 在普通组稀疏化[4−6] 中, 其罚项一般是对参数向量的 L_2 范数再取 L_1 范数. 目前, 学者们提出了罚项的多种范数组合, 罚项具有 L_q 范数形式[21] 的组稀疏化学习的目标函数为

$$\begin{aligned}\min_{\omega,b}-\sum_{m=1}^{n}\Bigg[\sum_{i=1}^{p}\Bigg[\log\phi_i\Big(x_i^m,\boldsymbol{b}_i\Big)+\sum_{j=i+1}^{p}\log\phi_{ij}(x_i^m,\\x_j^m,\boldsymbol{\omega}_{ij}\Big)\Bigg]\Bigg]+n\log Z(\omega,b)+\lambda\sum_{i=1}^{p}\sum_{j=i+1}^{p}\left\|\boldsymbol{\omega}_{ij}\right\|_q\end{aligned}\tag{10}$$

其中, $\lambda>0$ 为正则化参数, 若 $\lambda=0$ 即为普通两两关系马尔科夫网. 若 $q=2$, 上式即为普通组稀疏化模型. 当 $q=2$ 以及每个参数向量 $\boldsymbol{\omega}_{ij}$ 只包含单个变量时, 上式即为普通 L_1 正则化模型. 普通组稀疏模型中, 参数向量的 L_2 范数的 L_1 范数正则化, 可理解为参数向量 $\boldsymbol{\omega}_{ij}$ 的长度的 L_1 范数. 因此, 普通组稀疏化的目的是使得 $\boldsymbol{\omega}_{ij}$ 的长度实现稀疏性, 当向量长度为 0 时, 整个组的取值就为 0. L_2 范数没有方向上的侧重, 不会使得参数向量产生特殊的结构. L_1 范数则会对参数向量内部实现稀疏性, 而 L_∞ 范数则会使得参数向量内部的元素具有相同的大小. 本节主要利用参数向量的 L_2 范数的 L_1 正则化实现无向图的自适应组稀疏化学习.

为了避免对模较大的参数向量的有偏估计, 在式 (10) 的罚项中为每个参数向量分配不同的权值, 使参数向量的惩罚程度具有自适应性质[13]. 那么, 本节讨论的两两关系马尔科夫网的自适应组稀疏化学习的目标函数为

$$\begin{aligned}\min_{\omega,b}-\sum_{m=1}^{n}\Bigg[\sum_{i=1}^{p}\Bigg[\log\phi_i(x_i^m,\boldsymbol{b}_i)+\sum_{j=i+1}^{p}\log\phi_{ij}(x_i^m,\\x_j^m,\boldsymbol{\omega}_{ij})\Bigg]\Bigg]+n\log Z(\omega,b)+\lambda\sum_{i=1}^{p}\sum_{j=i+1}^{p}d_{ij}\beta_{ij}^2\left\|\boldsymbol{\omega}_{ij}\right\|_2\end{aligned}\tag{11}$$

其中, d_{ij} 为 $\boldsymbol{\omega}_{ij}$ 的长度, 即参数个数, $\beta_{ij}=1/\|\hat{\boldsymbol{\omega}}_{ij}\|_2$ 表示参数向量 $\boldsymbol{\omega}_{ij}$ 的权, 而 $\hat{\boldsymbol{\omega}}_{ij}$ 为式 (10) $q=2$ 时的普通组稀疏的估计值. 显然, 自适应组稀疏可以看作是两步的普通组稀疏. 第一步, 计算普通组稀疏的

估计值, 为了符号的简洁, 把负对数似然函数表示为 $L(\omega, b)$, 那么普通组稀疏的目标函数为

$$f_1(\omega, b, \lambda_1) = L(\omega, b) + \lambda_1 \sum_{i=1}^{p} \sum_{j=i+1}^{p} \|\boldsymbol{\omega}_{ij}\|_2 \quad (12)$$

普通组稀疏的估计值为 $\hat{\boldsymbol{\omega}} \equiv \hat{\boldsymbol{\omega}}(\lambda_1) = \arg\min_{\boldsymbol{\omega},\boldsymbol{b}} f_1(\boldsymbol{\omega}, b, \lambda_1)$. 第二步, 利用普通组稀疏的估计值得到罚函数的权

$$\beta_{ij} = \begin{cases} \dfrac{1}{\|\hat{\boldsymbol{\omega}}_{ij}\|_2}, & \|\hat{\boldsymbol{\omega}}_{ij}\|_2 > 0 \\ \infty, & \|\hat{\boldsymbol{\omega}}_{ij}\|_2 = 0 \end{cases} \quad (13)$$

那么自适应组稀疏化的目标函数为

$$f_2(\boldsymbol{\omega}, b, \lambda_2) = L(\boldsymbol{\omega}, b) + \lambda_2 \sum_{i=1}^{p} \sum_{j=i+1}^{p} d_{ij}\beta_{ij}^2 \|\boldsymbol{\omega}_{ij}\|_2 \quad (14)$$

这里定义 $0 \cdot \infty = 0$, 因此第一步实现的稀疏性并不会影响第二步产生的稀疏性. 自适应组稀疏的估计值为 $\hat{\boldsymbol{\omega}} \equiv \hat{\boldsymbol{\omega}}(\lambda_2) = \arg\min_{\boldsymbol{\omega},b} f_1(\omega, b, \lambda_2)$. 利用参数向量 $\boldsymbol{\omega}_{ij}$ 的普通组稀疏的解对其进行加权修正, 使模较大的参数向量具有较小的惩罚权值, 而使模较小的参数向量具有较大的惩罚权值, 实现了罚项的自适应调整.

Zhu 等提出一种 Grafting-light 算法[22], 该算法能够有效解决马尔科夫随机场中 L_1 范数的特征选择和结构学习问题, 总是能够得到全局最优解, 并选择出有效的重要特征. 该算法比 Grafting 算法表现更好, 未来我们可以考虑 Grafting-light 算法在自适应稀疏中的应用. Lee 等曾提出一种自适应的多任务套索算法[23], 对参数矩阵进行稀疏学习, 对其权值进行最大后验学习 (Maximum a posteriori, MAP), 通过轮转寻优法, 将参数向量 $\boldsymbol{\omega}_{ij}$ 的学习和参数向量的权值 β_{ij} 的学习分解为两个子优化问题进行求解. 本文的自适应组稀疏权值 β_{ij} 利用组稀疏的估计值 $\hat{\beta}_{ij}$, 简单直接, 避免了多任务问题的轮转求解过程.

2 目标函数的近似

本节讨论的通过稀疏模型生成的两两关系马尔科夫网节点之间的连接可能很稠密, 而归一化常量的计算耦合了网络的所有变量, 节点连接越稠密则其计算越难以处理. 因此, 必须通过对优化目标函数进行近似简化, 特别是归一化常量的近似简化, 能有效降低计算成本, 实现优化问题的可处理性.

2.1 伪似然近似

无向图模型中, 用于解决似然函数难以计算问题的一个近似办法是, 把似然函数替换为单变量条件概率分布的乘积. 该方法也称为伪似然 (Pseudo-likelihood) 近似法[15−17]. 无向图模型的伪似然函数近似法保持了解的一致性, 随着训练样本数的增加, 极大伪似然估计值收敛于极大似然估计值. 利用伪似然函数法, 原问题的似然函数改写为

$$L_{pl}(\omega, b) = -\sum_{m=1}^{n} \left[\sum_{i=1}^{p} \log p\left(x_i^m | \boldsymbol{X}_{-i}^m, \omega, b\right) \right] \quad (15)$$

$\boldsymbol{X}_{-i}$ 表示除变量 X_i 之外的其他所有变量. 上式的条件概率分布为已知变量 X_i 之外的其他所有变量时 X_i 的条件概率分布, 其具体形式如下:

$$p(x_i | \boldsymbol{X}_{-i}, \omega, b) = \frac{1}{Z_i} \phi_i(x_i, \boldsymbol{b}_i) \prod_{\{j|j\neq i\}} \phi_{ij}(x_i, x_j, \boldsymbol{\omega}_{ij}) \quad (16)$$

原来式 (2) 中需要求解全局归一化常量 $Z(\omega, b)$, 而利用伪似然函数法后, 只需求解局部归一化常量 Z_i, 而且 x_i 只需对 Z_i 的可能取值进行求和, 使得优化问题更易于求解. 经过伪似然近似后的自适应组稀疏化目标函数如下:

$$\min_{\omega,b} \; -\sum_{m=1}^{n} \left[\sum_{i=1}^{p} \left[\log \phi_i(x_i^m, \boldsymbol{b}_i) + \sum_{j=i+1}^{p} \log \phi_{ij}(x_i^m, x_j^m, \boldsymbol{\omega}_{ij}) + \log Z_i \right] \right] + \lambda \sum_{i=1}^{p} \sum_{j=i+1}^{p} d_{ij}\beta_{ij}^2 \|\boldsymbol{\omega}_{ij}\|_2 \quad (17)$$

2.2 变分近似

除伪似然近似法之外, 还可对归一化常量进行变分近似, 即对数划分函数的变分近似. 文献 [24−25] 对变分近似推理的方法进行了概述. Wainwright 等讨论了概率图模型的变分近似推理[20]. Lee 等提出了对数划分函数的 Bethe 自由能近似[12, 19, 26], 通过环信任传播算法[19] 实现. 平均场自由能近似[27] 也是图模型中较为经典的变分近似方法, 假设所有变量两两独立, 因此概率分布可表示为独立边缘分布的乘积.

所有的图模型都定义了一个联合概率分布 $p(\boldsymbol{x})$, 若存在某一个其他的近似联合概率分布 $b(\boldsymbol{x})$, $p(\boldsymbol{x})$ 和 $b(\boldsymbol{x})$ 之间的 "距离" 可使用 KL (Kullback-Leibler) 距离函数来衡量[28]:

$$D(b(\boldsymbol{x}) || p(\boldsymbol{x})) = \sum_{\boldsymbol{x}} b(\boldsymbol{x}) \ln \frac{b(\boldsymbol{x})}{p(\boldsymbol{x})} \quad (18)$$

由于 KL 距离是非负的, 当且仅当 $b(\boldsymbol{x})$ 和 $p(\boldsymbol{x})$ 相等时它才为 0, 所以 KL 距离十分适用于图模型中.

图模型中, "能量" $E(\boldsymbol{x})$ 的定义为: $p(\boldsymbol{x}) = \frac{1}{Z}\mathrm{e}^{-E(\boldsymbol{x})/T}$, 温度参数 T 只是一个定义能量单位量程的参数, 出于表达简洁性的考虑, 令 $T=1$. $p(\boldsymbol{x})$ 代入 KL 距离中, 有:

$$D(b(\boldsymbol{x})\,||p(\boldsymbol{x})) = \sum_{\boldsymbol{x}} b(\boldsymbol{x})\,E(\boldsymbol{x}) + \sum_{\boldsymbol{x}} b(\boldsymbol{x})\ln b(\boldsymbol{x}) + \ln Z \tag{19}$$

上式中的前两项一般定义为能量泛函:

$$F(b(\boldsymbol{x})) = \sum_{\boldsymbol{x}} b(\boldsymbol{x})\,E(\boldsymbol{x}) + \sum_{\boldsymbol{x}} b(\boldsymbol{x})\ln b(\boldsymbol{x}) = U(b(\boldsymbol{x})) + S(b(\boldsymbol{x})) \tag{20}$$

只有当能量泛函取得最小值 $F^* = -\ln Z$ 时, KL 距离为 0, 即近似概率分布等价于精确概率分布 $p(\boldsymbol{x})$. 由 KL 距离非负, 有 $F(b(\boldsymbol{x})) + \ln Z \geq 0$, 因此 $\ln Z \geq -F(b(\boldsymbol{x}))$, 即能量泛函能给出对数划分函数的下界. 能量泛函的第一项为平均能量, 第二项为熵 S 的负数.

1) 平均场自由能近似

在两两关系马尔科夫网的平均场自由能近似中, 近似联合概率分布可分解为每个节点概率分布的乘积, 其形式为

$$b(\boldsymbol{x}) = \prod_i b_i(x_i) \tag{21}$$

其中, $b_i(x_i)$ 满足条件 $\sum_i b_i(x_i) = 1$. 在平均场近似中, 单节点的概率分布为 $b_i(x_i)$, 而两节点的概率分布为 $b_{ij}(x_i,x_j) = b_i(x_i)\,b_j(x_j)$. 利用式 (21) 的近似联合概率分布函数, 能十分方便地计算出近似能量泛函. 两两关系马尔科夫网的能量函数为

$$E(\boldsymbol{x}) = -\sum_i \ln\phi_i(x_i) - \sum_{(ij)} \ln\phi_{ij}(x_i,x_j) \tag{22}$$

因此, 平均能量为

$$U_{MF}(\boldsymbol{b}) = -\sum_i\sum_{x_i} b_i(x_i)\ln\phi_i(x_i) - \sum_{(ij)}\sum_{x_i,x_j} b_i(x_i)\,b_j(x_j)\ln\phi_{ij}(x_i,x_j) \tag{23}$$

熵为

$$S_{MF}(\boldsymbol{b}) = -\sum_i\sum_{x_i} b_i(x_i)\ln b_i(x_i) \tag{24}$$

平均场能量泛函为 $F_{MF} = U_{MF} - S_{MF}$. 需寻找 b_i 的一个合适配置, 使得能量泛函取最小值, 即可得对数划分函数的一个近似.

对目标函数进行平均场自由能近似得到目标函数为

$$\min_{\omega,b} \; -\sum_{m=1}^{n}\Big[\sum_{i=1}^{p}\Big[\log\phi_i(x_i^m,\boldsymbol{b}_i) + \sum_{j=i+1}^{p}\log\phi_{ij}(x_i^m, x_j^m,\boldsymbol{\omega}_{ij})\Big]\Big] + n\log Z + \lambda\sum_{i=1}^{p}\sum_{j=i+1}^{p} d_{ij}\beta_{ij}^2\,\|\boldsymbol{\omega}_{ij}\|_2 \tag{25}$$

其中归一化常量的对数为

$$\log Z = -\sum_i\sum_{x_i} b_i(x_i)\ln\phi_i(x_i) - \sum_{(ij)}\sum_{x_ix_j} b_i(x_i)\,b_i(x_i)\ln\phi_{ij}(x_i,x_j) - \sum_{(i)}\sum_{x_i} b_i(x_i)\ln b_i(x_i) \tag{26}$$

2) Bethe 自由能近似

Bethe 自由能理论[26] 与平均场自由能理论在某些方面十分类似. Bethe 自由能中, 把能量泛函 F 表示为单节点概率分布 $b_i(x_i)$ 和两节点概率分布 $b_{ij}(x_i,x_j)$ 的一个函数. 上述概率分布需满足归一化条件 $\sum_i b_i(x_i) = \sum_{(ij)} b_{ij}(x_i,x_j) = 1$ 和边缘条件 $b_i(x_i) = \sum_j b_{ij}(x_i,x_j)$. 两两关系马尔科夫网只考虑单节点势能 $\phi_i(x_i)$ 和两节点势能 $\phi_{ij}(x_i,x_j)$, 对任意一个具有单节点概率分布 $b_i(x_i)$ 和两节点概率分布 $b_{ij}(x_i,x_j)$ 的近似联合概率分布来说, 平均能量为 $U = -\sum_i\sum_{x_i} b_i(x_i)\ln\phi_i(x_i) - \sum_{(ij)}\sum_{x_ix_j} b_{ij}(x_i,x_j)\ln\phi_{ij}(x_i,x_j)$.

精确的联合概率分布可写成如下形式

$$b(\boldsymbol{x}) = \frac{\prod\limits_{(ij)} b_{ij}(x_i,x_j)}{\prod\limits_i b_i(x_i)^{q_i-1}} \tag{27}$$

其中, q_i 为节点 x_i 的邻节点数. 根据信任传播规则, 节点 x_i 传播至 x_j 的信息包括除 x_j 外的所有邻节点的信息, 因此有 $q_i - 1$ 个. 利用式 (27), 可得熵 S 的 Bethe 近似形式:

$$S_{\text{Bethe}} = -\sum_{(ij)}\sum_{x_i,x_j} b_{ij}(x_i,x_j)\ln b_{ij}(x_i,x_j) + \sum_i (q_i-1)\sum_{x_i} b_i(x_i)\ln b_i(x_i) \tag{28}$$

对归一化常量进行 Bethe 自由能近似后, 目标

函数如下:

$$\min_{\omega,b} \; -\sum_{m=1}^{n}\Big[\sum_{i=1}^{p}\Big[\log\phi_i\left(x_i^m,\boldsymbol{b}_i\right)+\sum_{j=i+1}^{p}\log\phi_{ij}(x_i^m,x_j^m,\boldsymbol{\omega}_{ij})\Big]\Big]+n\log Z+\lambda\sum_{i=1}^{p}\sum_{j=i+1}^{p}d_{ij}\beta_{ij}^2\left\|\boldsymbol{\omega}_{ij}\right\|_2 \tag{29}$$

其中归一化常量的对数为

$$\begin{aligned}\log Z=&-\sum_i\sum_{x_i}b_i\left(x_i\right)\ln\phi_i\left(x_i\right)-\\&\sum_{(ij)}\sum_{x_ix_j}b_{ij}\left(x_i,x_j\right)\ln\phi_{ij}\left(x_i,x_j\right)-\\&\sum_{(ij)}\sum_{x_i,x_j}b_{ij}\left(x_i,x_j\right)\ln b_{ij}\left(x_i,x_j\right)+\\&\sum_i\left(q_i-1\right)\sum_{x_i}b_i\left(x_i\right)\ln b_i\left(x_i\right)\end{aligned} \tag{30}$$

3 精确目标函数的优化

由于自适应组稀疏化实际上是把普通组稀疏的估计值作为初始值, 对权值作出调整后再进行普通组稀疏的过程, 因此, 普通组稀疏的优化算法完全可适用于自适应组稀疏的求解. 目前, 普通组稀疏的优化算法一般有组最小角回归算法 (Group least angle regression, Group LARS)[29]、投影梯度算法、谱投影梯度算法 (Spectral projected gradient, SPG)[30−31]、块坐标下降算法[32]、投影拟牛顿法 (Projected quasi-Newton, PQN)[33]、轮换方向乘子法 (Alternating direction method of multipliers, ADMM)[34] 等. 本节主要讨论自适应组稀疏的 SPG 算法和 PQN 算法

3.1 谱投影梯度算法

在普通组稀疏的目标函数中, 由于正则化项的不可微, 考虑对其进行平滑处理. 对每一组 ij 引入附加变量 r_{ij}, 把正则化项中的每一个范数 $\left\|\boldsymbol{\omega}_{ij}\right\|_2$ 都替换为 r_{ij}, 并使得优化问题满足约束 $r_{ij}\geq\left\|\boldsymbol{\omega}_{ij}\right\|_2$. 因此, 平滑后的普通组稀疏的优化问题为

$$\begin{aligned}\min_{\omega,b,r}\quad & L\left(\omega,\boldsymbol{b}\right)+\lambda\sum_{i=1}^{p}\sum_{j=i+1}^{p}r_{ij}\\ \text{s.t.}\quad & r_{ij}\geq\left\|\boldsymbol{\omega}_{ij}\right\|_2,\quad\forall i,j\end{aligned} \tag{31}$$

上式把原来非线性不可微的正则化项替换为一个简单的线性函数. 对任意的可行对 $\{\omega,r\}$ 来说, 问题 (31) 给出了原问题的一个上界, 只有当所有组都满足 $r_{ij}=\left\|\boldsymbol{\omega}_{ij}\right\|_2$ 时问题 (31) 才取得最小值.

算法 1 (谱投影梯度算法).

输入: 目标函数 $f\left(\boldsymbol{x}\right)$, 投影函数 $\Pi_{\mathbf{C}}\left(\boldsymbol{x}\right)$, 初始参数向量 $\boldsymbol{x}_0$, 最优容许度 ε, 需要存储的前几步的函数个数 m, 充分下降参数 ν, 直线搜索的参数 ξ_1 和 ξ_2, 步长上限和下限 $\alpha_{\max}$ 和 $\alpha_{\min}$

$k\leftarrow 0$; $\boldsymbol{x}_0\leftarrow\Pi_{\mathbf{C}}\left(\boldsymbol{x}_0\right)$;//投影初始参数向量;

$\boldsymbol{f}_k\leftarrow f\left(\boldsymbol{x}_0\right)$;//计算目标函数

$\boldsymbol{g}_k\leftarrow\nabla f\left(\boldsymbol{x}_0\right)$;//计算梯度

while $\left\|\boldsymbol{x}_k-\Pi_{\mathbf{C}}\left(\boldsymbol{x}_k-\boldsymbol{g}_k\right)\right\|_\infty>\varepsilon$

do if $k=0$

then $\alpha\leftarrow-\min\left(1,1/\left\|\boldsymbol{g}_k\right\|_1\right)$;

//初始化步长

else $\alpha\leftarrow\boldsymbol{y}_k^{\mathrm{T}}\boldsymbol{s}_k/\boldsymbol{y}_k^{\mathrm{T}}\boldsymbol{y}_k$;//Barzilai-Borwein 步长

$\alpha\leftarrow\max\left(\alpha_{\min},\min\left(\alpha_{\max},\alpha\right)\right)$;//安全 BB 步长

$\boldsymbol{x}_{k+1}\leftarrow\Pi_{\mathbf{C}}\left(\boldsymbol{x}_k-\alpha\boldsymbol{g}_k\right)$;

$\boldsymbol{f}_{k+1}\leftarrow f\left(\boldsymbol{x}_{k+1}\right)$;

$\boldsymbol{g}_{k+1}\leftarrow\nabla f\left(\boldsymbol{x}_{k+1}\right)$;

while $f_{k+1}>\max_{i=k-m:k}f_i+\nu\boldsymbol{g}_k^{\mathrm{T}}\left(\boldsymbol{x}_{k+1}-\boldsymbol{x}_k\right)$;

do select $\alpha\in\left(\xi_1\alpha,\xi_2\alpha\right)$;

$\boldsymbol{g}_{k+1}\leftarrow\nabla f\left(\boldsymbol{x}_{k+1}\right)$; $\boldsymbol{s}_k\leftarrow\boldsymbol{x}_{k+1}-\boldsymbol{x}_k$;

$\boldsymbol{y}_k\leftarrow\boldsymbol{g}_{k+1}-\boldsymbol{g}_k$; $k\leftarrow k+1$;

投影梯度法是求解约束优化问题的算法, 其优化问题为

$$\min_{\boldsymbol{x}\in\mathbf{C}}\quad f\left(\boldsymbol{x}\right) \tag{32}$$

其中, $f\left(\boldsymbol{x}\right)$ 为可微函数, 而 $\mathbf{C}$ 为闭凸集. 这里的优化变量可认为是问题 (31) 中的变量对 $(\omega,\boldsymbol{r})$, 而闭凸集即为问题 (31) 的约束条件定义的集合. 投影梯度算法的迭代形式为

$$\boldsymbol{x}_{k+1}\leftarrow\Pi_{\mathbf{C}}\left(\boldsymbol{x}_k-\alpha\nabla f\left(\boldsymbol{x}_k\right)\right) \tag{33}$$

步长 α 满足回溯线搜索的 Armijo 条件, 而 $\Pi_{\mathbf{C}}$ 为闭凸集 $\mathbf{C}$ 上的欧氏投影算子, 其定义如下

$$\Pi_{\mathbf{C}}\left(\boldsymbol{x}\right)=\arg\min_{\boldsymbol{y}\in\mathbf{C}}\left\|\boldsymbol{x}-\boldsymbol{y}\right\|_2 \tag{34}$$

投影梯度算法存在两方面的问题: 1) 投影步骤的计算过于复杂; 2) 最速下降步长的使用会导致收敛速度过慢. 文献 [11] 提出的 SPG 算法, 对投影梯度算法进行两方面的修改. 一是线搜索的步长改为

$$\alpha_{bb}=\frac{\boldsymbol{y}_k^{\mathrm{T}}\boldsymbol{s}_k}{\boldsymbol{y}_k^{\mathrm{T}}\boldsymbol{y}_k} \tag{35}$$

其中, $\boldsymbol{y}_k=\nabla f\left(\boldsymbol{x}_{k+1}\right)-\nabla f\left(\boldsymbol{x}_k\right)$, $\boldsymbol{s}_k=\boldsymbol{x}_{k+1}-\boldsymbol{x}_k$. 该步长也称为 Barzilai-Borwein 步长. 二是使用非单调搜索技术, 要求使最近的某些次迭代的目标函

数减小即可. 非单调的 Armijo 条件如下

$$f\left(\boldsymbol{x}_{k+1}\right) \leq \max_{i=k-m:k} f\left(\boldsymbol{x}_i\right)+\nu \nabla f\left(\boldsymbol{x}_k\right)^{\mathrm{T}}\left(\boldsymbol{x}_{k+1}-\boldsymbol{x}_k\right) \tag{36}$$

其中, $\nu \in (0,1)$. 非单调的 Armijo 条件可能会接受某些使目标函数上升的步长 α_{bb}, 但是它能保持全局收敛性. 式 (36) 中所考虑的前几步的函数值的个数一般设置为 10. 算法 1 给出了谱投影梯度算法的伪代码.

问题 (31) 中, 每一个约束只作用于与组 ij 相关的参数向量 $\boldsymbol{\omega}_{ij}$, 因此可通过计算每一个组的投影问题, 从而得到所有组的投影. 使用 SPG 算法, 不仅显著减少迭代次数, 同时还能保证约束集上的投影计算十分有效.

定理 1. 对问题 (31) 的每一个组来说, 投影问题为

$$\begin{aligned} \Pi_{\mathbf{C}_2}\left(\boldsymbol{\omega}_{ij}, r_{ij}\right)=\arg \min_{\boldsymbol{y}_{ij} \in \mathbf{C}_2, z_{ij}} & \left\|\begin{bmatrix}\boldsymbol{\omega}_{ij} \\ r_{ij}\end{bmatrix}-\begin{bmatrix}\boldsymbol{y}_{ij} \\ z_{ij}\end{bmatrix}\right\|_2 \\ \text{s.t.} \quad & z_{ij} \geq\left\|\boldsymbol{y}_{ij}\right\|_2 \end{aligned} \tag{37}$$

投影问题 (37) 的解为

$$\Pi_{\mathbf{C}_2}\left(\boldsymbol{\omega}_{ij}, r_{ij}\right)=\begin{cases} \left(\boldsymbol{\omega}_{ij}, r_{ij}\right), & \\ \left(\frac{\boldsymbol{\omega}_{ij}}{\left\|\boldsymbol{\omega}_{ij}\right\|_2} \frac{\left\|\boldsymbol{\omega}_{ij}\right\|_2+r_{ij}}{2}, \frac{\left\|\boldsymbol{\omega}_{ij}\right\|_2+r_{ij}}{2}\right), & \\ (0,0), & \end{cases}$$

$$\begin{aligned} &\left\|\boldsymbol{\omega}_{ij}\right\|_2 \leq r_{ij} \\ &\left\|\boldsymbol{\omega}_{ij}\right\|_2 > r_{ij}, \quad \left\|\boldsymbol{\omega}_{ij}\right\|_2+r_{ij}>0 \\ &\left\|\boldsymbol{\omega}_{ij}\right\|_2 > r_{ij}, \quad \left\|\boldsymbol{\omega}_{ij}\right\|_2+r_{ij} \leq 0 \end{aligned} \tag{38}$$

求解时可以利用范数的非负性, 可把问题 (37) 等价表示为

$$\begin{aligned} &\arg \min_{\boldsymbol{y}_{ij}, z_{ij}} \frac{1}{2}\left\|\boldsymbol{\omega}_{ij}-\boldsymbol{y}_{ij}\right\|_2^2+\frac{1}{2}\left(r_{ij}-z_{ij}\right)^2 \\ &\text{s.t.} \quad z_{ij} \geq\left\|\boldsymbol{y}_{ij}\right\|_2 \end{aligned} \tag{39}$$

详细求解过程见文献 [4].

3.2 投影拟牛顿算法

虽然 SPG 算法是目前求解组稀疏问题较为有效的方法, 但是当目标函数比较复杂时, SPG 算法的计算成本非常高. 由于普通组稀疏问题转换为可微的约束优化问题, 如上节给出的问题 (31). 针对具有这种结构的组稀疏问题, Lee 等提出有限制内存 PQN 算法[12], 对 Hessian 阵作出近似, 能够处理高维约束优化问题. PQN 算法的目标问题是在凸集 $\mathbf{C}$ 上最小化目标函数 $f\left(\boldsymbol{x}\right)$.

PQN 算法的每一步迭代都对目标函数在当前迭代值 $\boldsymbol{x}_k$ 附近作二次近似:

$$\begin{aligned} q_k\left(\boldsymbol{x}\right)=f\left(\boldsymbol{x}_k\right)+\left(\boldsymbol{x}-\boldsymbol{x}_k\right)^{\mathrm{T}} \nabla f\left(\boldsymbol{x}_k\right)+ \\ \frac{1}{2}\left(\boldsymbol{x}-\boldsymbol{x}_k\right)^{\mathrm{T}} B_k\left(\boldsymbol{x}-\boldsymbol{x}_k\right) \end{aligned} \tag{40}$$

其中, B_k 是 Hessian 阵的一个正定近似. 为了得到可行的下降方向, PQN 算法在凸集 $f\left(\boldsymbol{x}\right)$ 上寻找式 (40) 二次近似的最小值 $\boldsymbol{x}_k^*$, 即

$$\boldsymbol{x}_k^*=\arg \min_{\boldsymbol{x} \in \mathbf{C}} \quad q_k\left(\boldsymbol{x}\right) \tag{41}$$

那么, 下降方向为 $\boldsymbol{d}=\boldsymbol{x}_k^*-\boldsymbol{x}_k$, 其中, 当 $\alpha \in [0,1]$ 时 $\alpha \in [0,1]$ 是可行的. 该下降方向可作为回溯线搜索的方向, 直到新的迭代值满足 Armijo 条件时停止. 当式 (40) 中的 B_k 为精确 Hessian 阵时, 通常先取 $\alpha=1$. PQN 算法在满足二阶充分条件的最小值邻域内具有二次收敛速率. 但是, 这种方法具有两方面的缺点: 1) 需要计算稠密的 $n \times n$ 维近似 Hessian 阵; 2) 约束二次模型的最小值的求解可能具有很高的计算成本.

算法 2 (投影拟牛顿算法).

输入: 目标函数 $f\left(\boldsymbol{x}\right)$, 投影函数 $\Pi_{\mathbf{C}}\left(\boldsymbol{x}\right)$, 初始参数向量 $\boldsymbol{x}_0$, 最优容许度 ε, 修正个数 m, 充分下降参数 ν, 直线搜索参数 ξ_1 和 ξ_2, SPG 迭代上限 c.

$k \leftarrow 0$; $\boldsymbol{x}_0 \leftarrow \Pi_{\mathbf{C}}\left(\boldsymbol{x}_0\right)$;
$\boldsymbol{f}_k \leftarrow f\left(\boldsymbol{x}_0\right)$; $\boldsymbol{g}_k \leftarrow \nabla f\left(\boldsymbol{x}_0\right)$;
while $\left\|\boldsymbol{x}_k-\Pi_{\mathbf{C}}\left(\boldsymbol{x}_k-\boldsymbol{g}_k\right)\right\|_{\infty}>\varepsilon$
do $\alpha=1$;
if $k=0$ then
$\boldsymbol{d}_k \leftarrow -\boldsymbol{g}_k \min\left(1,1/\left\|\boldsymbol{g}_k\right\|_1\right)$;
else
$\boldsymbol{x}_k^*=SPG\left(\boldsymbol{x}_k, c, \boldsymbol{g}_k, \sigma, \boldsymbol{S}, \boldsymbol{Y}\right)$;
$\boldsymbol{d}_k=\boldsymbol{x}_k^*-\boldsymbol{x}_k$; $\boldsymbol{x}_{k+1} \leftarrow \boldsymbol{x}_k+\alpha \boldsymbol{d}_k$;
$\boldsymbol{g}_k \leftarrow \nabla f\left(\boldsymbol{x}_0\right)$; $\boldsymbol{g}_{k+1} \leftarrow \nabla f\left(\boldsymbol{x}_{k+1}\right)$;
while $f_{k+1}>f_k+\nu \boldsymbol{g}_k^{\mathrm{T}}\left(\boldsymbol{x}_{k+1}-\boldsymbol{x}_k\right)$ do
select $\alpha \in\left(\xi_1 \alpha, \xi_2 \alpha\right)$;
$\boldsymbol{x}_{k+1} \leftarrow \boldsymbol{x}_k+\alpha \boldsymbol{d}_k$;
$\boldsymbol{f}_{k+1} \leftarrow f\left(\boldsymbol{x}_{k+1}\right)$; $\boldsymbol{g}_{k+1} \leftarrow \nabla f\left(\boldsymbol{x}_{k+1}\right)$;
$\boldsymbol{s}_k \leftarrow \boldsymbol{x}_{k+1}-\boldsymbol{x}_k$; $\boldsymbol{y}_k \leftarrow \boldsymbol{g}_{k+1}-\boldsymbol{g}_k$;
if $k>m$ then
把 $\boldsymbol{S}$ 和 $\boldsymbol{Y}$ 中最旧的变量删掉;
$\boldsymbol{S} \leftarrow [\boldsymbol{S} \quad \boldsymbol{s}_k]$
$\boldsymbol{Y} \leftarrow [\boldsymbol{Y} \quad \boldsymbol{y}_k]$
$\sigma \leftarrow \left(\boldsymbol{y}_k^{\mathrm{T}} \boldsymbol{s}_k\right)/\left(\boldsymbol{y}_k^{\mathrm{T}} \boldsymbol{y}_k\right)$
$k \leftarrow k+1$;

关于 Hessian 阵的近似策略, 拟牛顿法一般先取一个比例单位矩阵 $B_0=\sigma I$ 作为 Hessian 阵的近似, 然后 B_{k+1} 的每一步更新都需要参数变化和梯度

变化满足割线方程:

$$B_{k+1}\boldsymbol{s}_k = \boldsymbol{y}_k \tag{42}$$

其中, $\boldsymbol{s}_k = \boldsymbol{x}_{k+1} - \boldsymbol{x}_k$, $\boldsymbol{y}_k = \nabla f(\boldsymbol{x}_{k+1}) - \nabla f(\boldsymbol{x}_k)$. 式 (42) 中 B_{k+1} 的解不是唯一的, 可通过 BFGS 准则确定唯一解. 而 Kovács 等提出了有限内存的 BFGS (Broyden-Fletcher-Gold farb-Shannon) 更新算法 (L-BFGS)[11], 并不存储 B_{k+1} 的值, 而是存储 m 个 $\boldsymbol{s}_k$ 和 $\boldsymbol{y}_k$ 的集合.

给出 B_k 的 L-BFGS 表示之后, 可利用 SPG 算法求解问题 (41). 在 $q_k(\boldsymbol{x})$ 及其梯度的计算过程中, SPG 的运行成本主要来源于谱投影 $\Pi_{\mathbf{C}}$ 的计算. 本文讨论的无向图自适应组稀疏化问题中, 投影的计算只需线性时间成本, 而目标函数的计算却是难的, 近似目标函数的计算成本也比投影计算成本高得多. 所以, PQN 能有效处理无向图自适应组稀疏化问题.

算法 2 给出了 PQN 算法的伪代码. 伪代码中的 $SPG(\boldsymbol{x}_k, c, \boldsymbol{g}_k, \sigma, \boldsymbol{S}, \boldsymbol{Y})$ 表示从 $\boldsymbol{x}_k$ 开始通过 SPG 的 c 次迭代求解问题 (41), 梯度设为 $\boldsymbol{g}_k$, 而 L-BFGS 表示的构造使用参数 σ、$\boldsymbol{S}$ 和 $\boldsymbol{Y}$.

4 实验结果

4.1 实验数据介绍及预处理

本实验使用一个人工合成数据集和 5 个 UCI (University of Californina Irvine) 数据集进行具有不同边势形式的两两关系马尔科夫网的结构学习问题, 还对目标函数作出不同的近似, 以及采用不同的优化算法求解得到的实验结果作了比较. 在实验过程中, 目标模型的正则化参数选择为 $\lambda = 2^x$, 其中 x 的取值范围为从 10 下降到 −3, 下降步长为 0.25. 样本数据均分为两部分, 一部分作为训练数据集, 用于训练模型参数, 另一部分作为测试数据集, 用于模型性能评估.

人工数据集的合成首先随机生成一个具有 10 个节点的两两关系马尔科夫网, 每个节点有 3 种取值状态, 其单节点势能和两节点势能的参数 $\boldsymbol{b}_i$ 和 $\boldsymbol{\omega}_{ij}$ 为服从标准正态分布的随机数. 然后从该网络中随机产生 4 000 个样本, 均分为训练数据集和测试数据集.

用户知识建模数据集, 共 403 个实例, 来自 http://archive.ics.uci.edu/ml/datasets/User+Knowledge+Modeling. 数据集包括 6 个属性, 其中 5 个为输入值, 1 个为目标值. 去除结论性属性 6, 原数据的其他 5 个属性的取值是 0 到 1, 即相应比值, 所以, 属性值取 1 (实际值为 0 到 0.2)、取 2 (实际值为 0.2 到 0.4)、取 3 (实际值为 0.4 到 0.6)、取 4 (实际值为 0.6 到 0.8)、取 5 (实际值为 0.8 到 1). 最终用于实验的数据为一个 403×5 的矩阵, 且每个元素有 5 种取值状态.

定性破产数据集, 共 250 例, 来自 http://archive.ics.uci.edu/ml/datasets/Qualitative_Bankruptcy. 数据集共 7 个属性, 最后一个属性为结论. 去除结论属性 7, 其他属性分别用 1、2、3 代替 P、A、N. 最终用于实验的数据为一个 250×6 的矩阵, 且每个元素有三种取值状态.

威斯康辛州乳腺癌数据集, 来自 http://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Original). 总共 699 个样例, 每个样本有 11 个属性. 首先, 属性 1 是一个编号, 属性 11 是最终结论, 它们都不是病征属性, 因此去掉属性 1 和属性 11, 保留属性 2 到属性 10. 处理后剩下 9 个属性, 且每一个属性有 10 种取值情况. 又由于每个属性的取值状态数较多, 所以对取值状态再进行以下处理: 取值为 1 (当实际值是 1、2 或 3 时), 取值为 2 (当实际值为 4、5 或 6 时), 取值为 3 (当实际值为 7、8、9 或 10 时), 这样处理基本保证了均匀分布, 所以处理后的数据与原数据的概率分布基本不变. 最终用于实验的数据为一个 699×9 的矩阵, 且每个元素有三种取值状态.

急性炎症数据集, 共 120 个样本, 来自 http://archive.ics.uci.edu/ml/datasets/Acute+Inflammations. 数据集共 8 个属性.

体温高于 38°C 赋值为 2, 低于 38°C 的为 1, 其余属性的 No 为 1, Yes 用 2 代替. 因此最后剩余 8 个属性, 用于实验的是 120×8 的矩阵, 每个元素有 2 种取值状态.

汽车评估数据集, 共 1 728 个样本, 来自 http://archive.ics.uci.edu/ml/datasets/Car+Evaluation. 每个样本数据集共 6 个属性.

数据中的 Low 和 Med 分别用 1, 2 代替, High 和 Vhigh 用 3 代替, Small 和 Big 用 1 和 3 代替, 车门数 2 用 1 代替, 3, 4 用 3 代替, 5more 用 3 代替, 处理后用于实验的数据集为 1 728×6 的矩阵, 每个元素有 3 种取值状态. 天气建模数据集, 共有 540 个样本, 来自 http://archive.ics.uci.edu/ml/datasets/Climate+Model+Simulation+Crashes, 每个样本包含 21 个属性, 其中属性 1、属性 2 和属性 21 分别为: 研究序号、模拟序号和模拟结果, 去除这 3 个属性, 剩余 18 个属性. 处理后用于实验的数据集为 540×18 的矩阵, 每个元素有 2 种取值状态.

4.2 实验结果与分析

两两关系马尔科夫网在多值变量情况下的结构学习, 其边势函数可分为三种形式, 即对角相同边势、对角不同边势和完全边势. 这里, 首先比较三种

不同边势形式对模型性能的影响. 实验过程中, 由于对角相同边势类似于二值变量的稀疏学习情况, 因此无须使用组稀疏化, 只使用一般的 L_1 正则化项.

而对角不同边势和完全边势由于边势函数中具有多个参数, 可进行组稀疏化学习, 因此对于这两种边势本实验分别使用一般 L_1 正则化、组稀疏化 (算法用 Group 表示) 和自适应组稀疏化算法 (算法用 Agroup 表示) 进行结构学习. 另外, 目标函数全部使用精确的目标函数, 而优化算法使用投影拟牛顿算法.

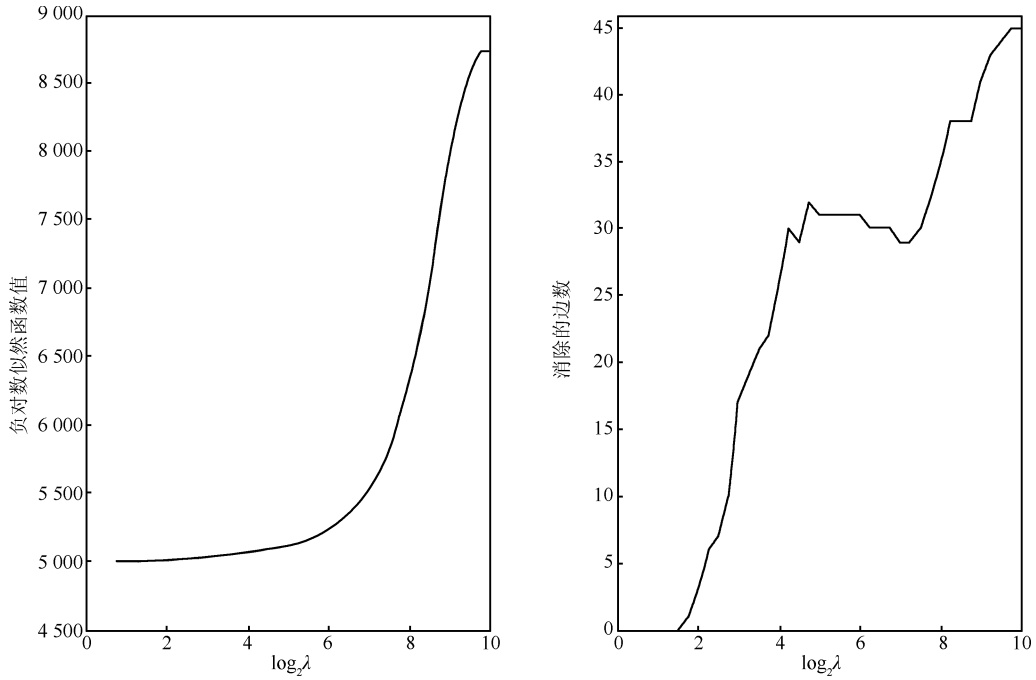
模型的拟合程度使用负对数似然函数 (Negative log-likelihood, NLL) 值进行比较. NLL 值越小, 说明模型对当前数据的拟合程度越好. 而模型的稀疏程度用被消除的边数 (Number of eliminated edges, NOE) 来表示. NOE 值越大, 被消除的边越多, 即模型更简单. 模型设定越复杂, 对当前数据的拟合程度越好, 但过于复杂, 花费太大, 且有可能使得模型只适用于当前样本集, 造成过拟合.

我们在人工合成数据集上进行实验, 以完全边势情况下的组稀疏为例, 观察正则化参数 λ 对模型拟合能力 (负对数似然函数值) 和稀疏性能 (消除边数) 的影响. 因不同边势和 L_1 正则化项与组稀疏的实验结果随着正则化参数 λ 变化而变化的趋势类似, 此处只给出一组数据的实验结果.

图 1 (a) 给出了在人工合成数据集上使用完全边势进行组稀疏化学习实验的结果. 从图 1 (a) 可以看出, 虽然实验结果有小的波动, 总体的趋势是: 正则化参数 λ 很小时, NOE 为 0, 随着 λ 增大, NLL 值越大, NOE 值也越大. 最后 NLL 值和 NOE 值均趋于稳定. 这是因为惩罚力度越大, 模型稀疏性越强, 模型也越简单, 但同时模型对数据的拟合程度也变差, 所以并不是 λ 越大, 模型越简单越好. 因此, 正则化参数 λ 选择时, 既要让 NLL 值尽可能的小, 保证模型对数据的拟合程度, 也要使 NOE 值尽可能大, 使模型更加简单.

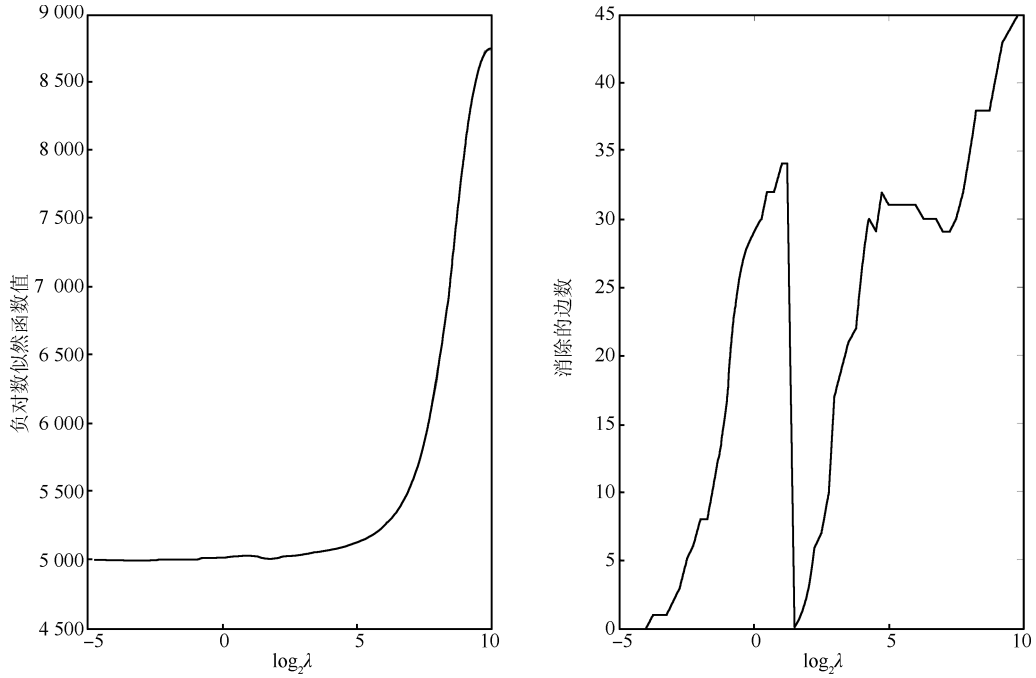
图 1 (b) 给出了在人工合成数据集上使用完全边势进行自适应组稀疏化学习的实验结果. 对比图 1 (a) 和图 1 (b) 可以看出, 在 λ 较小时, 普通的组稀疏算法没有稀疏效果, 而自适应组稀疏算法能够使模型稀疏化, 且效果非常明显. 我们对不同数据集进行实验, 并取定性破产数据集的一组实验结果进行详细对比.

图 2 给出了定性破产数据集在使用三种不同边势和不同正则化模型时的实验结果. 从图 2 可以看出, L_1 正则化模型的拟合能力最差, 组稀疏和自适应组稀疏均优于 L_1 稀疏, 且自适应组稀疏的拟合能力最优, 其中使用完全边势的自适应组稀疏的拟合能力最好. L_1 正则化模型的运行时间最短, 自适应组稀疏的运行时间最长, 约是组稀疏的两倍. L_1 正则化的运行时间比组稀疏和自适应组稀疏短, 这是因为 L_1 正则化的参数向量只包含单个变量. 自适应组稀疏可以理解为两步组稀疏, 第一步组稀疏得到罚函数的权值, 第二步利用上一步得到的权值再进行稀疏, 因此运行时间翻倍. 图 3 给出了定性破产



(a) 人工数据集上使用完全边势进行组稀疏在不同 λ 值下的实验结果

(a) The experimental results on synthetic dataset using group regularization and full potentials under different λ values



(b) 人工数据集上使用完全边势进行自适应组稀疏时在不同 λ 值下的实验结果

(b) The experimental results on synthetic dataset using adaptive group regularization and full potentials under different λ values

图 1 人工数据集上使用完全边势进行不同稀疏学习时在不同 λ 值下的实验结果

Fig. 1 The experimental results on synthetic dataset using different regularizations and full potentials under different λ values

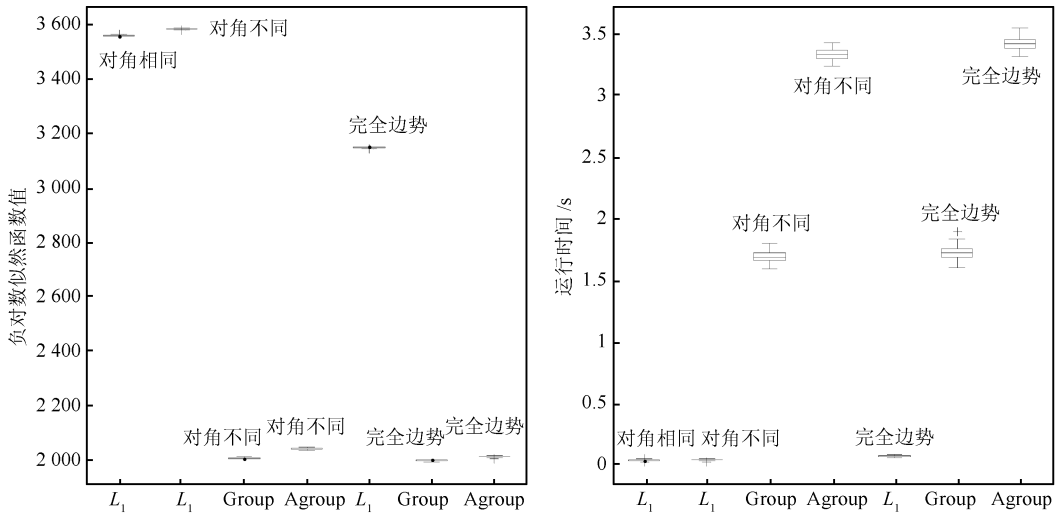


图 2 定性破产数据集上使用不同边势进行不同稀疏学习算法时的实验结果

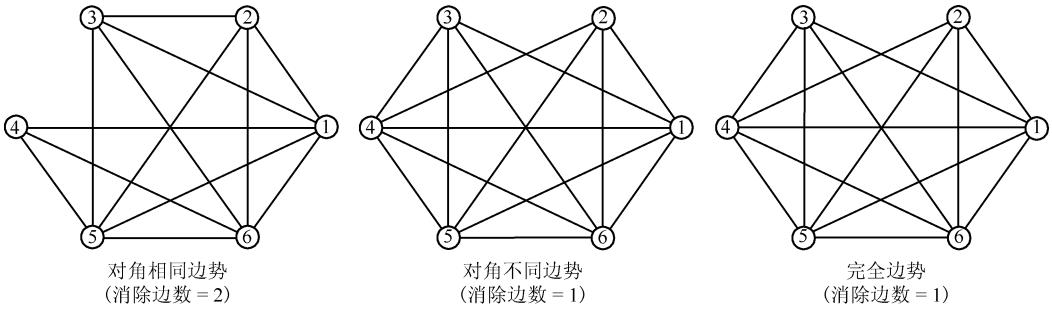
Fig. 2 The experimental results on qualitative bankruptcy dataset using different regularizations and potentials

数据集在不同边势情况下的不同正则化模型的网络图. 从图 3 可以看出, 自适应组稀疏算法的稀疏化性能最好, L_1 正则化模型次之. L_1 正则化模型的稀疏性能优于组稀疏是因为, L_1 正则化消除一条边只需要使单个变量为 0, 组稀疏需要使一组变量为 0, 组稀疏更复杂. 综合来看, 完全边势情况下的自适应组稀疏模型具有最好的性能.

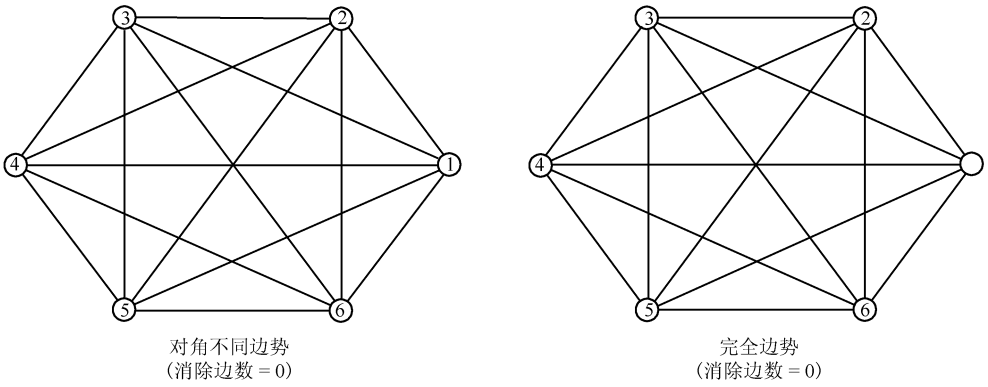
图 3 中两条边的权值分别为: $w_{1,2} = 1.1463$, $w_{1,6} = 0.7977$. 可以看出属性 1 行业风险最重要, 属性 2 管理风险次之, 属性 6 操作风险再次.

表 1 和表 2 分别给出了包括人工数据集在内的多个数据集在完全边势情况下进行不同正则化

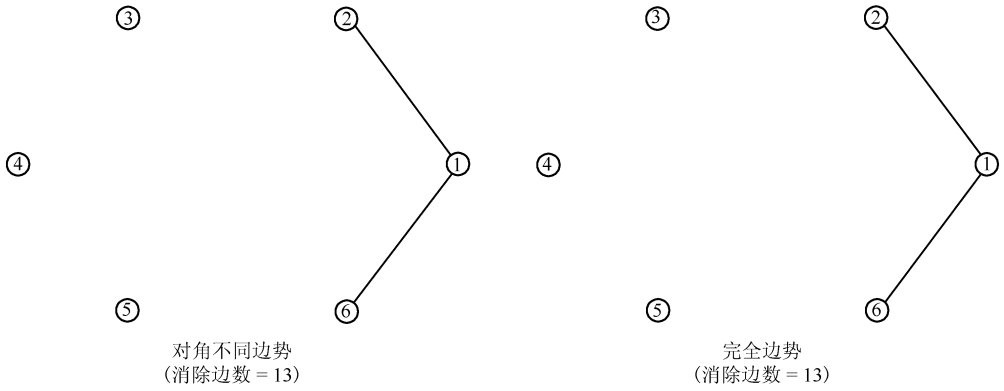
的实验结果及未使用稀疏算法的普通两两马尔科夫网的 NLL 值和 NOE 值. 我们可以看到自适应组稀疏始终具有比 L_1 正则化和组稀疏化学习更好的稀疏性能. 自适应组稀疏要比 L_1 稀疏耗费更多的时间, 但在大多数数据集上自适应组稀疏的 NLL 值和 NOE 值均比 L_1 稀疏更好, 即自适应组稀疏牺牲计算成本, 使其拟合性能和稀疏性能均比 L_1 好.



(a) 定性破产数据集上使用不同边势进行 L_1 稀疏学习时的网络图
(a) Netplots of qualitative bankruptcy dataset using L_1 regularization and different potentials



(b) 定性破产数据集上使用不同边势进行组稀疏学习时的网络图
(b) Netplots of qualitative bankruptcy dataset using group regularization and different potentials



(c) 定性破产数据集上使用不同边势进行自适应稀疏学习时的网络图
(c) Netplots of qualitative bankruptcy dataset using adaptive group regularization and different potentials

图 3　定性破产数据集上使用不同边势进行不同稀疏学习算法时的网络图

Fig. 3　Netplots of qualitative bankruptcy dataset using different regularizations and different potentials

表 1 多个数据集使用完全边势进行不同稀疏化学习时的实验结果

Table 1 The experimental results using full potential and different regularizations

数据集	Number of original edges	Number of eliminated edges			Running time (s)		
		L_1	Group	Agroup	L_1	Group	Agroup
人工数据集	45	17	0	34	32.622	24.996	54.706
威斯康辛州乳腺癌	36	2	0	25	2.376	8.489	12.890
用户知识建模	10	0	0	9	0.110	2.422	4.129
定性破产	15	2	0	13	0.035	1.724	3.423
急性炎症	28	0	0	18	0.128	1.769	3.522
汽车评估	15	0	0	14	0.171	1.71	1.789

表 2 多个数据集使用完全边势进行不同稀疏化学习模型和原始两两 Markov 模型的 NLL 值对比

Table 2 The comparison NLL results between different regularized models using full potentials and original pairwise Markov model

数据集	Negative loglikelihood of original pairwise Markov network			Test set negative loglikelihood of sparse pairwise Markov network		
	L_1	Group	Agroup	L_1	Group	Agroup
人工数据集	5 146.842	4 987.297	4 986.097	5 227.051	4 997.738	5 022.105
威斯康辛州乳腺癌	1 378.134	1 378.027	1 378.105	1 362.493	1 363.614	1 526.246
用户知识建模	3 944.299	2 259.696	2 390.159	1 910.785	1 754.665	1 698.807
定性破产	3 140.56	2 074.585	2 215.701	3 559.324	1 995.566	2 010.704
急性炎症	6 120.487	1 822.075	1 951.788	1 637.608	1 691.303	1 607.565
汽车评估	5 596.330	5 596.331	5 596.331	5 595.209	5 595.761	5 543.82

4.3 三种近似目标函数

本节通过实验比较目标函数的三种近似方法的性能, 包括伪似然近似、平均场自由能近似和 Bethe 自由能近似, 这三种近似方法分别简化表示为 Pseudo、Mean 和 Bethe. 其中, 精确目标函数和伪似然近似为严格凸函数, 而变分近似中的平均场自由能近似和 Bethe 自由能近似为非凸函数. 对三种近似目标函数分别进行 L_1 正则化、组稀疏正则化和自适应组稀疏正则化学习, 优化算法使用投影拟牛顿算法.

图 4 给出了在天气建模数据集上使用精确目标函数和三种近似目标函数进行不同正则化的实验结果. 图 5 给出了天气建模数据集上使用精确目标函数和三种近似目标函数的不同稀疏化学习模型的网络图. 我们可以看出, 三种近似目标函数的运行时间远远比精确目标函数小, 伪似然近似函数的拟合性能略优于精确目标函数, 稀疏性能也比精确目标函数略好. 自适应组稀疏的拟合性能优于其他两种模型, 且稀疏效果优势明显.

表 3～ 5 分别给出了使用三种不同近似目标函数在多个数据集上进行不同稀疏化学习的实验结果. 对比表 3～5, 我们发现使用精确目标函数的模型在大多数数据集上的稀疏性能略好于伪似然近似, 但伪似然近似目标函数在某些数据集上的稀疏性能优于精确目标函数, 但两者的差异不大. 近似目标函数的 NLL 值比精确目标函数低. 平均场近似稀疏性能的拟合性能和稀疏性能都是最差.

综合看来, 伪似然近似与精确目标函数在小数据集上表现相当, 但在大数据上, 伪似然近似的优势非常明显, 运行时间大大减少, 性能最次是平均场近似. 无论使用哪一种目标函数, 自适应组稀疏的稀疏性能优势明显.

4.4 两种优化算法

本节实验比较精确目标函数在不同正则化稀疏学习情况中使用两种不同优化算法的性能, 这两种优化算法为第 3 节中介绍的谱投影梯度算法和投影拟牛顿算法, 分别表示为 spg 和 pqn. 下面的实验中使用完全边势, 目标函数使用精确目标函数 (近似目标函数得到类似结果, 这里不再讨论).

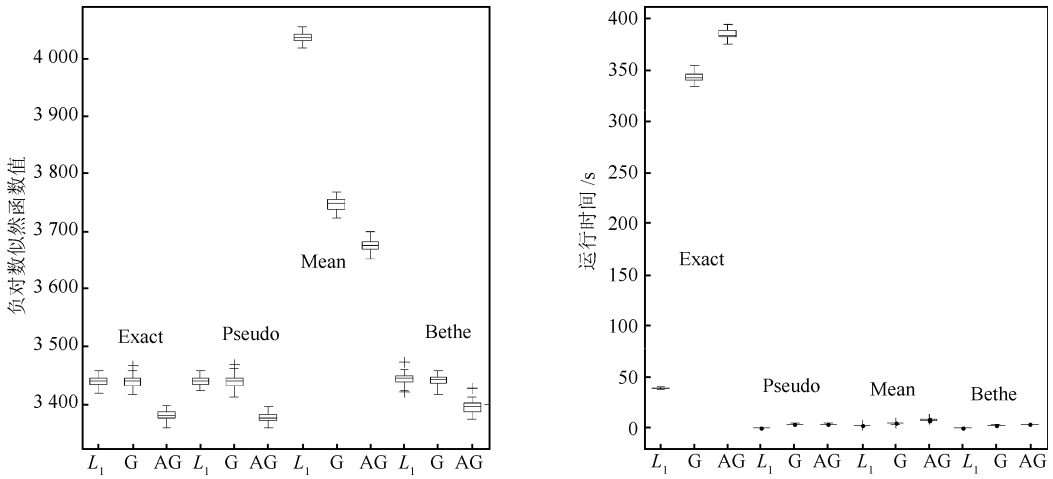
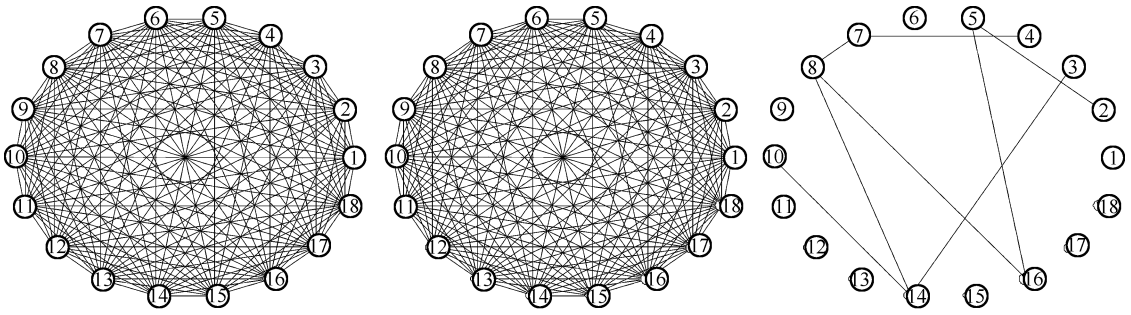


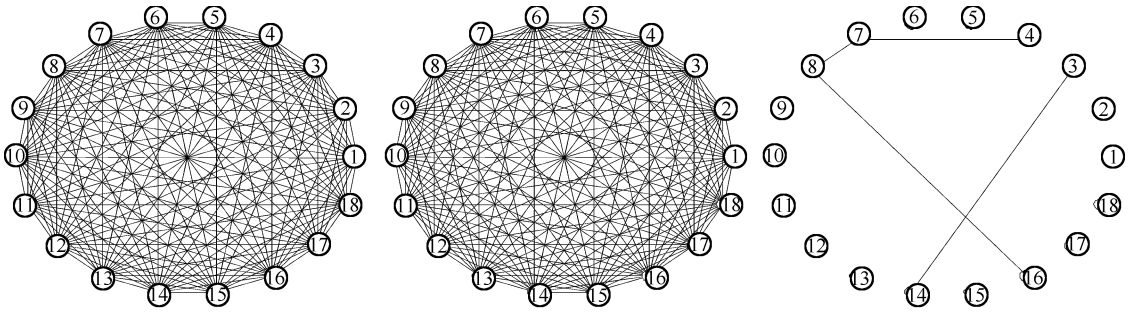
图 4　天气建模数据集使用精确目标函数和三种近似函数进行不同正则化稀疏学习的实验结果

Fig. 4　The experimental results on climate model dataset using exact object function and three approximations and different regularizations



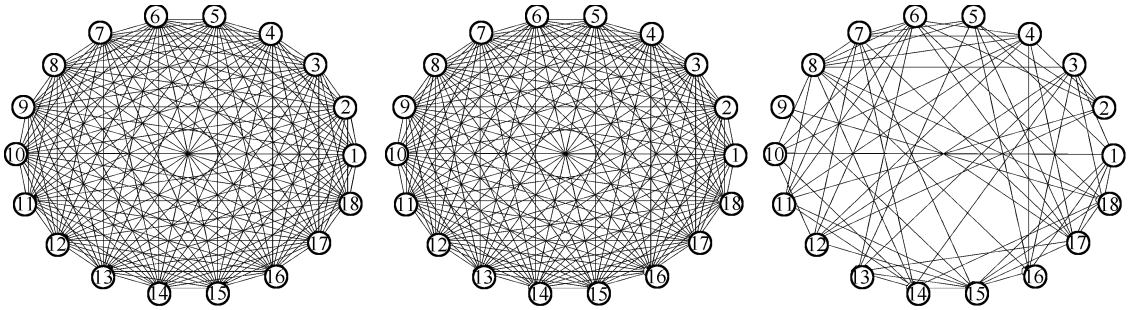
(a) 天气建模数据集使用精确目标函数的不同正则化稀疏学习的网络图

(a) Netplots of climate model dataset using exact object function and different regularizations

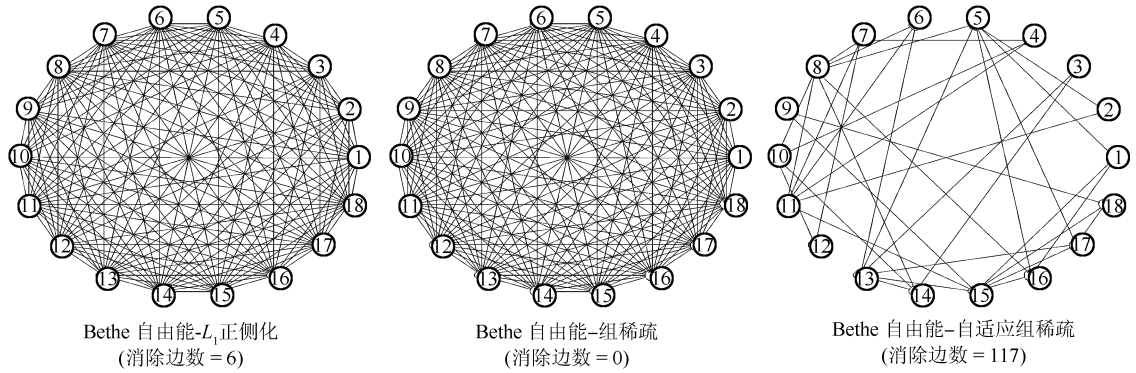


(b) 天气建模数据集使用伪似然近似目标函数的不同正则化稀疏学习的网络图

(b) Netplots of climate model dataset using pseudo approximation and different regularizations



(c) 天气建模数据集使用平均场自由能近似目标函数的不同正则化稀疏学习的网络图
(c) Netplots of climate model dataset using mean field approximation and different regularizations



(d) 天气建模数据集使用 Bethe 自由能近似目标函数的不同正则化稀疏学习的网络图
(d) Netplots of climate model dataset using bethe free energy approximation and different regularizations

图 5 天气建模数据集使用精确目标函数和不同近似目标函数的不同正则化稀疏学习的网络图
Fig. 5 Netplots of climate model dataset using exact object function and different approximations and different regularizations

表 3 多个数据集使用伪似然近似函数进行不同稀疏化学习时的实验结果
Table 3 The experimental results using pseudo approximation and different regularizations

数据集	Test set negative log-likelihood			Number of eliminated edges			Running time		
	L_1	Group	Agroup	L_1	Group	Agroup	L_1	Group	Agroup
人工数据集	5 002.630	4 991.167	4991.167	7	12	12	0.768	2.043	2.069
威斯康辛州乳腺癌	1 335.790	1 336.343	1487.128	0	0	19	0.226	1.953	2.692
用户知识建模	1 946.843	1 816.045	1710.943	0	0	9	0.188	2.113	4.185
定性破产	2 867.103	2 038.121	2001.807	0	0	10	0.110	1.767	3.599
急性炎症	1 755.517	1 669.925	1685.517	9	0	18	0.263	1.717	1.979
汽车评估	5 595.775	5 596.047	5549.359	0	0	12	0.622	2.628	2.755
天气建模数据集	3 438.427	3 439.168	3377.977	2	0	149	0.1824	3.9282	3.9754

图 6 给出了使用精确目标函数时, 投影拟牛顿算法和谱投影梯度算法在不同正则化项时的实验结果. 从图 6 可以看出使用 spg 优化算法所得到的 NLL 值更低, 但相差并不大. 自适应组稀疏比组稀疏的 NLL 值低, 即拟合性能更好. pqn 优化算法的运行时间比 spg 优化算法的运行时间更长, 自适应组稀疏的运行时间比组稀疏增加一倍. 本实验使用的是小数据, 因此运行时间差异并不大, 若用于大数据优化问题求解, spg 优化算法的运算成本优势就会变大. 图 7 给出了两种不同优化算法的不同正则化项稀疏学习的模型网络图. 从图 7 可以看出两种优化算法的稀疏性能相当, 自适应组稀疏的稀疏性能优势明显.

表 4　多个数据集使用平均场近似函数进行不同稀疏化学习时的实验结果

Table 4　The experimental results using mean field approximation and different regularizations

数据集	Test set negative log-likelihood			Number of eliminated edges			Running time		
	L_1	Group	Agroup	L_1	Group	Agroup	L_1	Group	Agroup
人工数据集	8 610.858	6 391.883	6 391.882	0	0	0	0.060	1.224	2.211
威斯康辛州乳腺癌	1 698.240	1 723.024	1 723.024	0	0	0	0.039	0.031	0.043
用户知识建模	1 591.604	1 603.425	1 613.418	0	0	0	0.039	0.032	0.048
定性破产	1 275.923	2 141.670	2 142.670	0	0	0	0.032	0.328	0.342
急性炎症	880.933	847.765	847.767	0	0	0	0.337	0.713	1.303
汽车评估	5 621.832	5 549.61	5 549.61	0	0	0	0.033	0.027	0.034
天气建模数据集	4 036.119	3 748.372	3 674.349	0	0	83	2.4663	4.4765	7.9722

表 5　多个数据集使用 Bethe 自由能近似函数进行不同稀疏化学习时的实验结果

Table 5　The experimental results using Bethe free energy approximation and different regularizations

数据集	Test set negative log-likelihood			Number of eliminated edges			Running time		
	L_1	Group	Agroup	L_1	Group	Agroup	L_1	Group	Agroup
人工数据集	7 639.921	8 030.764	11 594.352	0	0	0	0.061	0.058	0.987
威斯康辛州乳腺癌	1 685.804	1 729.666	1 729.666	0	0	0	0.044	0.025	0.036
用户知识建模	1 589.197	1 599.749	1 613.733	0	0	0	0.021	0.037	0.058
定性破产	1 249.866	1 359.077	1 359.077	0	0	0	0.015	0.015	0.023
急性炎症	3 486.903	1 019.181	1 019.181	9	0	0	0.359	0.234	0.510
汽车评估	5 618.367	5 548.880	5 548.880	0	0	0	0.017	0.022	0.030
天气建模数据集	3 440.048	3 440.141	3 397.049	6	0	117	0.2361	2.9716	3.4122

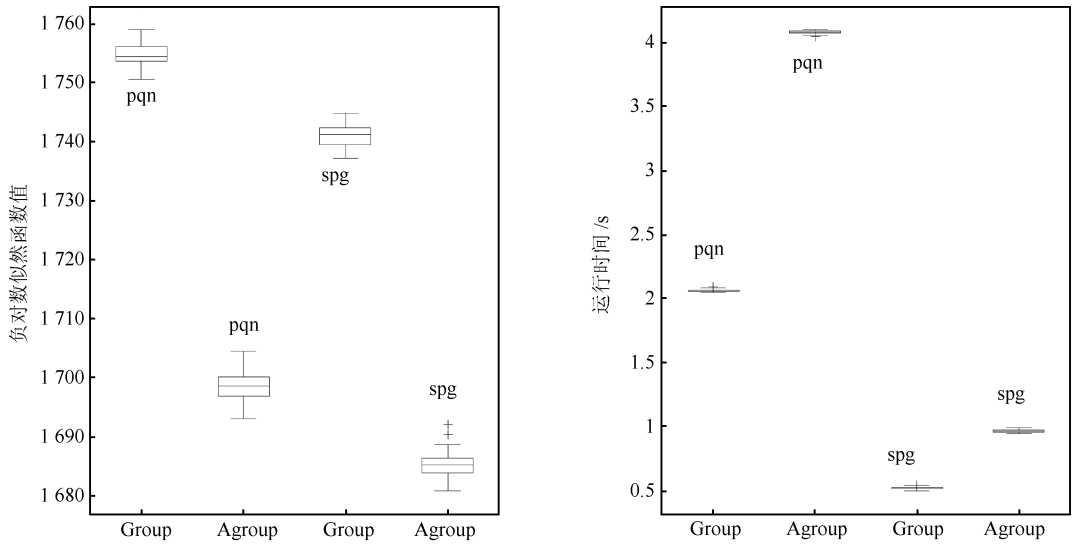
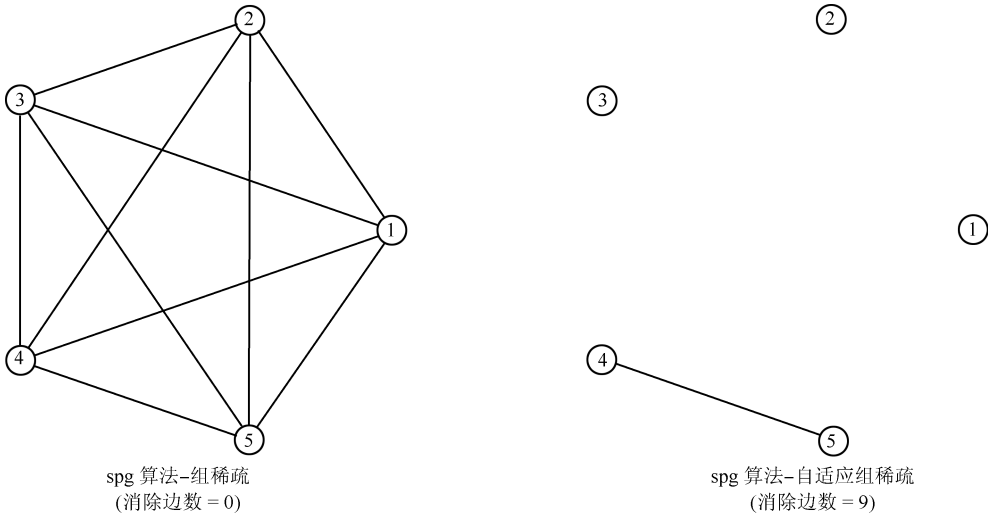
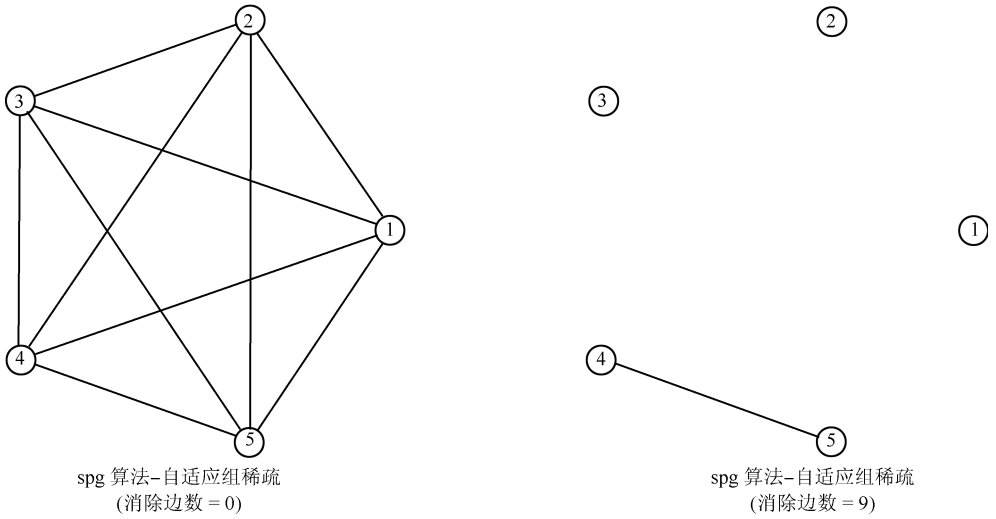


图 6　不同优化算法的不同正则化稀疏学习在用户知识建模数据上的实验结果

Fig. 6　The experimental results on user knowledge model dataset using different optimizations and regularizations



(a) 用户知识建模使用 SPG 算法进行不同正则化稀疏学习的模型网络图
(a) Netplots of user knowledge model dataset using SPG algorithm and different regularizations



(b) 用户知识建模使用 PQN 算法进行不同正则化稀疏学习的模型网络图
(b) Netplots of user knowledge model dataset using PQN algorithm and different regularizations

图 7 用户知识建模使用不同优化算法进行不同正则化稀疏学习的模型网络图

Fig. 7 Netplots of user knowledge model dataset using different algorithms and different regularizations

稀疏网络图中剩余的一条边的权值为: $\omega_{4,5} = 0.5094$, 属性 4 和属性 5 分别为 LPR (与目标科目相关科目的用户考试成绩) 和 PEG (目标科目的用户考试成绩). 即用户的知识水平与 PEG 和 LPR 密切相关. 表 6 给出了 spg 优化算法在多个数据集上进行不同稀疏学习的实验结果. 从表 6 可以看出自适应组稀疏的稀疏性能优势很明显, 且 NLL 值相差不大, 甚至低于组稀疏. 对比表 1, 我们发现 pqn 算法的稀疏性能比 spg 算法的稀疏性能更好. 综合来看, 无论是 spg 算法还是 pqn 算法, 自适应组稀疏的性能都更好.

5 总结和展望

在两两关系马尔科夫网的离散多值变量情况下, 每条边的参数不再为单一的某个值, 而是与之相关的一组参数, L_1 正则化不再适用于该场景, 因此本文对此给出两两关系马尔科夫网的组稀疏化学习, 并且为了避免估计有偏性, 还提出两两关系马尔科夫网的自适应组稀疏化学习. 实验结果表明, 自适应组稀疏不仅稀疏性能优于 L_1 正则化, 且拟合性能也比 L_1 正则化好.

为了使目标函数能应用在复杂网络中, 本文对目标函数进行三种近似, 分别为伪似然近似、平均场

表 6 多个数据集使用 spg 算法的不同稀疏化学习的实验结果

Table 6 Experiment results on different datasets using spg algorithm and different regularizations

数据集	Test set negative log-likelihood		Number of eliminated edges		Running time	
	Group	Agroup	Group	Agroup	Group	Agroup
人工数据集	5 024.690	5 043.908	0	32	23.987	47.801
威斯康辛州乳腺癌	1 362.947	1 527.148	0	24	6.363	12.744
用户知识建模	1 741.362	1 684.713	0	9	0.523	0.965
定性破产	1 819.702	1 880.150	0	13	0.275	0.476
急性炎症	1 412.900	1 283.992	0	15	0.223	0.376
汽车评估	5 595.836	5 545.103	0	14	0.274	0.255

自由能近似和 Bethe 自由能近似. 实验表明, 自适应组稀疏的性能在精确目标函数和伪似然函数中有很大的性能优势, 后两种近似函数的所有稀疏学习方法均未能稀疏成功. 精确目标函数的运行时间最长, 三种近似函数的运行时间大大减短. 不过精确目标函数的稀疏性能最好, 伪似然函数次之, 平均场近似和 Bethe 自由能近似的稀疏性能很差.

本文还给出了求解两两关系马尔科夫网自适应组稀疏化学习问题的两种优化算法 (谱投影梯度算法和投影拟牛顿算法). 实验表明 pqn 算法的稀疏性能比 spg 算法的稀疏性能更好, 但 spg 算法的运行时间更短. 无论使用哪一种优化算法, 自适应组稀疏的都具有更好的性能. 本文提出的自适应组稀疏模型在迭代过程中对参数进行组处理, 使模较小的参数向量有较大的惩罚, 从而剔除该组参数向量对应的边, 实现模型的稀疏化. 在 1 组人工数据集和 5 组 UCI 数据集上的实验表明, 本文的自适应组稀疏模型具有较好的性能.

References

1 Liu Jian-Wei, Li Hai-En, Luo Xiong-Lin. Learning technique of probabilistic graphical models: a review. *Acta Automatica Sinica*, 2014, **40**(6): 1025−1044
(刘建伟, 黎海恩, 罗雄麟. 概率图模型学习技术研究进展. 自动化学报, 2014, **40**(6): 1025−1044)

2 Liu Jian-Wei, Li Hai-En, Luo Xiong-Lin. Representation theory of probabilistic graphical models. *Computer Science*, 2014, **41**(9): 1−17
(刘建伟, 黎海恩, 罗雄麟. 概率图模型表示理论. 计算机科学, 2014, **41**(9): 1−17)

3 Liu Jian-Wei, Cui Li-Peng, Luo Xiong-Lin. Survey on the sparse learning of probabilistic graphical models. *Chinese Journal of Computers*, 2014, **37**: Online Publishing No. 114
(刘建伟, 崔立鹏, 罗雄麟. 概率图模型的稀疏化学习综述. 计算机学报, 2014, **37**: 在线出版号 No. 114)

4 Schmidt M. Graphical Model Structure Learning with L_1 Regularization [Ph. D. dissertation], The University of British Columbia, Vancouver, BC, 2010

5 Yuan M, Lin Y. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2006, **68**(1): 49−67

6 Huang J Z, Zhang T. The benefit of group sparsity. *The Annals of Statistics*, 2010, **38**(4): 1978−2004

7 Schlüter F. A survey on independence-based Markov networks learning. *Artificial Intelligence Review*, 2014, **42**(4): 1069−1093

8 Lee K, Anguelov D, Sumengen B, Gokturk S B. Markov random field models for hair and face segmentation. In: Proceedings of 8th IEEE International Conference on Automatic Face and Gesture Recognition. Amsterdam, The Netherlands: IEEE, 2008. 1−6

9 Amizadeh S, Hauskrecht M. Latent variable model for learning in pairwise Markov networks. In: Proceedings of the 2010 AAAI Conference on Artificial Intelligence. Atlanta, Georgia, USA: AAAI, 2010. 382−387

10 Zhang X H, Saha A, Vishwanathan S V N. Accelerated training of max-margin Markov networks with kernels. *Theoretical Computer Science*, 2014, **519**: 88−102

11 Kovács E, Szántai T. Discovering the Markov network structure [Online], available: http://www.arxiv.org, July 2, 2013

12 Lee S I, Ganapathi V, Koller D. Efficient structure learning of Markov networks using L_1-regularization. In: Proceedings of the 20th Annual Conference on Neural Information Processing Systems. Vancouver, British Columbia, Canada, 2007. 817−824

13 Wang H S, Leng C L. A note on adaptive group lasso. *Computational Statistics and Data Analysis*, 2008, **52**(12): 5277−5286

14 Wei F R, Huang J. Consistent group selection in high-dimensional linear regression. *Bernoulli: Official Journal of the Bernoulli Society for Mathematical Statistics and Probability*, 2010, **16**(4): 1369−1384

15 Besag J. Efficiency of pseudolikelihood estimation for simple Gaussian fields. *Biometrika*, 1977, **64**(3): 616−618

16 Kok S, Domingos P. Learning the structure of Markov logic networks. In: Proceedings of the 22nd International Conference on Machine Learning. Bonn, Germany: ACM, 2005. 441−448

17 Campbell N D F, Subr K, Kautz J. Fully-connected CRFs with non-parametric pairwise potential. In: Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Portland, OR, USA: IEEE, 2013. 1658−1665

18 Krähenbühl P, Koltun V. Efficient inference in fully connected CRFs with Gaussian edge potentials [Online], available: http://www.arxiv.org/abs/1210.5644, October 20, 2012

19 Yedidia J S, Freeman W T, Weiss Y. Constructing free-energy approximations and generalized belief propagation algorithms. *IEEE Transactions on Information Theory*, 2005, **51**(7): 2282−2312

20 Wainwright M J, Jordan M I. Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 2008, **1**(1−2): 1−305

21 Liu J, Ye J P. Efficient l_1/l_q norm regularization [Online], available: http://www.arxiv.org/abs/1009.4766v1, September 24, 2010

22 Zhu J, Lao N, Xing E P. Grafting-light: fast, incremental feature selection and structure learning of Markov random fields. In: Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Washington D. C., USA: ACM, 2010. 303−312

23 Lee S, Zhu J, Xing E P. Adaptive multi-task lasso: with application to eQTL detection. In: Proceedings of 24th Annual Conference on Neural Information Processing Systems. Vancouver, British Columbia, Canada, 2010. 1306−1314

24 Cheng Qiang, Chen Feng, Dong Jian-Wu, Xu Wen-Li. Variational approximate inference methods for graphical models. *Acta Automatica Sinica*, 2012, **38**(11): 1721−1734
(程强, 陈峰, 董建武, 徐文立. 概率图模型中的变分近似推理方法. 自动化学报, 2012, **38**(11): 1721−1734)

25 Li Hai-En, Liu Jian-Wei, Luo Xiong-Lin Variational approximate inference for probabilistic graphical models. In: Proceeding of the 2013 Chinese Intelligent Automation Conference (4). Yangzhou, Jiangsu, China, 2013.
(黎海恩, 刘建伟, 罗雄麟. 概率图模型的变分近似推理. 见: 2013 年中国智能自动化学术会议论文集 (第四分册). 扬州, 江苏, 中国, 2013.)

26 Yedidia J S, Freeman W T, Weiss Y. Constructing free-energy approximations and generalized belief propagation algorithms. *IEEE Transactions on Information Theory*, 2005, **51**(7): 2282−2312

27 Sun S L. A review of deterministic approximate inference techniques for Bayesian machine learning. *Neural Computing and Applications*, 2013, **23**(7−8): 2039−2050

28 Györfi L, Györfi Z, Vajda I. Bayesian decision with rejection. *Problems of Control and Information Theory*, 1979, **8**(5−6): 445−452

29 Yuan M, Lin Y. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2006, **68**(1): 49−67

30 Birgin E, Martínez J, Raydan M. Nonmonotone spectral projected gradient methods on convex sets. *SIAM Journal on Optimization*, 2000, **10**(4): 1196−1211

31 Yu Z S. Solving bound constrained optimization via a new nonmonotone spectral projected gradient method. *Applied Numerical Mathematics*, 2008, **58**(9): 1340−1348

32 Fu W J. Penalized regressions: the bridge versus the lasso. *Journal of Computational and Graphical Statistics*, 1998, **7**(3): 397−416

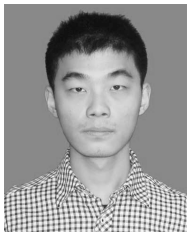
33 Schmidt M, van den Berg E, Friedlander M P, Murphy K. Optimizing costly functions with simple constraints: a limited-memory projected quasi-Newton algorithm. In: Proceedings of the 12th International Conference on Artificial Intelligence and Statistics. Clearwater Beach, Florida, USA, 2009. 456−463

34 Boyd S, Parikh N, Chu E, Peleato B, Eckstein J. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, 2011, **3**(1): 1−122

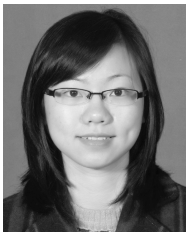
刘建伟 博士, 中国石油大学 (北京) 副研究员. 主要研究方向为稀疏学习, 智能信息处理, 复杂系统的分析. 预测与控制, 算法分析与设计. 本文通信作者.
E-mail: liujw@cup.edu.cn
(**LIU Jian-Wei** Ph. D., associate professor in the Department of Automation, College of Geophysics and Information Engineering, China University of Petroleum, Beijing Campus (CUP). His research interest covers sparse

learning, intelligent information processing, analysis, prediction, controlling of complicated nonlinear system, and analysis of the algorithm and the designing. Corresponding author of this paper.)

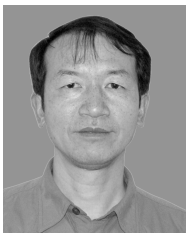
任正平 中国石油大学 (北京) 地球物理与信息工程学院硕士研究生. 主要研究方向为机器学习.
E-mail: renzhengping1225@sina.com
(**REN Zheng-Ping** Master student in the Department of Automation, College of Geophysics and Information Engineering, China University of Petroleum, Beijing Campus (CUP). Her main research interests is machine learning.)



刘泽宇 中国科学院软件研究所基础软件国家工程研究中心硕士研究生. 主要研究方向为机器学习.
E-mail: logonod@163.com
(**LIU Ze-Yu** Master student at the National Engineering Research Center for Fundamental Software, Institute of Software, Chinese Academy of Sciences. His main research interest is machine learning.)



黎海恩 中国石油大学 (北京) 地球物理与信息工程学院硕士研究生. 主要研究方向为机器学习, 概率图模型表示、学习和推理. E-mail: lihaien1988@163.com
(**LI Hai-En** Master student in the Department of Automation, College of Geophysics and Information Engineering, China University of Petroleum, Beijing Campus (CUP). Her main research interest is machine learning, probabilistic graphical model representation, learning and reasoning.)



罗雄麟 博士, 中国石油大学 (北京) 教授. 主要研究方向为智能控制, 复杂系统分析、预测与控制.
E-mail: luoxl@cup.edu.cn
(**LUO Xiong-Lin** Ph. D., professor in the Department of Automation, College of Geophysics and Information Engineering, China University of Petroleum, Beijing Campus (CUP). His research interest covers intelligent control, analysis, prediction, controlling of complicated nonlinear system.)