中国科学: 信息科学 2022 年 第 52 卷 第 1 期: 86–98
SCIENTIA SINICA Informationis
论文

面向半监督聚类的最优间隔分布学习机

张腾[1,2,3,4*], 黎铭[5], 金海[1,2,3,4]

1. 华中科技大学大数据技术与系统国家地方联合工程研究中心, 武汉 430074
2. 华中科技大学服务计算技术与系统教育部重点实验室, 武汉 430074
3. 华中科技大学集群与网格计算湖北省重点实验室, 武汉 430074
4. 华中科技大学计算机科学与技术学院, 武汉 430074
5. 南京大学计算机软件新技术国家重点实验室, 南京 210023
* 通信作者. E-mail: tengzhang@hust.edu.cn

收稿日期: 2021–05–30; 修回日期: 2021–08–24; 接受日期: 2021–09–01; 网络出版日期: 2022–01–05

国家自然科学基金 (批准号: 62006088, 62076121) 资助项目

摘要 基于间隔的聚类是一类经典的聚类算法, 此类算法假设聚类结构能通过引入监督学习中的间隔来确定. 即一个好的聚类结果, 当以其簇标记作为类别标记进行监督学习时, 所得分类器产生的关于间隔的目标物理量也同时达到最优. 目前最为有效的间隔物理量是间隔分布, 其基于最新的间隔理论, 取得了比优化最小间隔更好的效果. 然而在现实聚类任务中, 我们往往还能获得一些额外的监督信息, 例如两两样本之间的 "必连" 约束和 "勿连" 约束, 此时优化间隔分布是否还有效尚未可知. 对此, 本文提出面向半监督聚类的最优间隔分布学习机 (ODMSSC), 对该问题进行初步探索. ODMSSC 对应的形式化是一个混合整数规划, 我们将其放松成一个鞍点问题, 并提出一种高效的交替优化方法进行求解. 最终通过真实数据集上的实验, 我们验证了所提算法的有效性.

关键词 半监督聚类, 约束聚类, 最优间隔分布学习机, 间隔分布, 间隔

1 引言

聚类是机器学习中经典的无监督学习任务, 其基本想法是将数据按照相似度进行划分, 相似的样本在同一组, 不相似的样本在不同组. 在计算机的很多领域都有类似需求的任务, 如信息检索、计算机视觉、生物信息学等, 在过去的几十年间提出了形形色色的聚类算法 [1~5].

基于间隔 (margin) 的聚类是一类非常经典的算法 [6, 7], 此类算法假设聚类结构能通过引入监督学习中的间隔来确定. 即一个好的聚类结果, 当以其给出的簇标记作为类别标记进行监督学习时, 所得分类器产生的关于间隔的目标物理量也同时达到最优. 具体来说, 假设样本已经被赋予了簇标记, 将

引用格式: 张腾, 黎铭, 金海. 面向半监督聚类的最优间隔分布学习机. 中国科学: 信息科学, 2022, 52: 86–98, doi: 10.1360/SSI-2021-0187
Zhang T, Li M, Jin H. Optimal margin distribution machine for semi-supervised clustering (in Chinese). Sci Sin Inform, 2022, 52: 86–98, doi: 10.1360/SSI-2021-0187

此簇标记当作类别标记可以得到一个有监督信息的数据集, 在此数据集上以目标间隔物理量作为优化目标可以得到一个分类器, 最后再反过来将其预测结果作为对样本的簇标记赋值, 因此基于间隔的聚类往往被形式化成簇标记赋值和目标间隔物理量的联合优化. 根据所用间隔物理量的不同, 现有算法可以分为两类: 一类优化最小间隔 [8~10]; 另一类 [11] 基于最新的间隔理论 [12], 以间隔分布为优化目标, 取得了比前者更好的性能.

然而在现实聚类任务中, 除了大量的无标记样本, 我们往往还能获得一些额外的监督信息 [13], 利用这些信息进一步提升算法的聚类性能称为半监督聚类 (semi-supervised clustering, SSC). 监督信息大致有两种类型: 第 1 种是两两样本之间的成对约束信息, 包括 "必连" 约束和 "勿连" 约束, 前者是指两个样本必属于同一个簇, 后者是指两个样本必不属于同一个簇; 第 2 种则是少量的有标记样本, 注意有标记样本可以导出 "必连" 约束和 "勿连" 约束, 而反过来不成立, 因此第 1 种监督信息的形式更为一般化, 其对应的算法也更为通用.

尽管在聚类问题上, 优化间隔分布的有效性已被证明 [11], 但对于半监督聚类, 结论尚未可知. 本文提出面向半监督聚类的最优间隔分布学习机 (optimal margin distribution machine for semi-supervised clustering, ODMSSC) 对此进行初步的探索. 该方法沿袭了基于间隔的聚类方法的基本思路, 联合优化簇标记赋值和间隔分布, 但在优化簇标记赋值时通过利用监督信息, 既提升了优化求解的速度, 又改进了算法的聚类性能. ODMSSC 对应的形式化是一个混合整数规划问题, 难以直接求解, 我们将其放松成一个凸的鞍点问题, 并提出一种高效的交替优化方法进行求解. 最终通过真实数据集上的实验, 我们验证了所提算法的有效性.

本文第 2 节概述相关工作. 第 3 节详细介绍 ODMSSC 算法. 第 4 节给出相应的优化过程和实现细节. 第 5 节通过在真实数据集上的实验验证所提算法的有效性. 最后第 6 节总结全文, 并对未来工作做出展望.

2 相关工作

为后文叙述方便, 本节先给出一些符号约定, 记 $\mathcal{X} \subseteq \mathbb{R}^d$ 为样本空间, $\mathcal{Y} = \{1, -1\}$ 为类别标记空间, $\mathcal{D}$ 是定义在 $\mathcal{X} \times \mathcal{Y}$ 上的未知概率分布, $\mathcal{S} = \{(\boldsymbol{x}_i, y_i)\}_{i\in[m]}$ 是独立同分布 (i.i.d.) 地采样于分布 $\mathcal{D}$ 的数据集, 其中 $[m]$ 表示整数集合 $\{1, 2, \ldots, m\}$. $\phi : \mathcal{X} \mapsto \mathbb{H}$ 是正定核函数 κ 对应的特征映射, 即 $\kappa(\boldsymbol{x}, \boldsymbol{z}) = \langle\phi(\boldsymbol{x}), \phi(\boldsymbol{z})\rangle$. $[\boldsymbol{A}]_{ij}$ 表示矩阵 $\boldsymbol{A}$ 第 i 行第 j 列的元素. $\mathrm{diag}(\boldsymbol{y})$ 表示以 $\boldsymbol{y}$ 为主对角线的方阵. $\mathrm{sgn}(\cdot)$ 是符号函数, 当输入为正、负时分别输出 1 和 -1. $1_{\{\cdot\}}$ 是指示函数, 当输入为真时输出 1, 否则输出 0.

2.1 最优间隔分布学习机

假设空间由线性模型 $h(\boldsymbol{x}) = \langle\boldsymbol{w}, \phi(\boldsymbol{x})\rangle$ 构成, 最终模型预测的类别标记为 $\mathrm{sgn}(h(\boldsymbol{x}))$. 有标记样本 $(\boldsymbol{x}, y)$ 的间隔 [14] 是其到分类超平面的有向距离 $\gamma_h(\boldsymbol{x}, y) = yh(\boldsymbol{x}) = y\langle\boldsymbol{w}, \phi(\boldsymbol{x})\rangle$.

支持向量机 (support vector machine, SVM) 优化的是最小间隔 [15,16], 即样本到分类超平面的最短距离, 因此所得分类超平面只由少数支持向量构成, 其余非支持向量全被忽略了. 该方法对噪声异常敏感, 当支持向量的类别标记存在噪声时, 只能得到次优的分类效果 [17].

受间隔理论 [12] 的启发, 最优间隔分布学习机 (optimal margin distribution machine, ODM) 采用间隔分布的一阶统计量和二阶统计量来刻画整个分布, 要求间隔均值最大化的同时间隔方差最小化 [18],

于是可得如下形式化:

$$\min_{\boldsymbol{w},\bar{\gamma},\xi_i,\epsilon_i} \frac{1}{2}\|\boldsymbol{w}\|^2 - \eta\bar{\gamma} + \frac{\lambda}{2m}\sum_{i\in[m]}(\xi_i^2+\epsilon_i^2),$$
$$\text{s.t. } \gamma_h(\boldsymbol{x}_i,y_i) \geqslant \bar{\gamma}-\xi_i,\ \gamma_h(\boldsymbol{x}_i,y_i) \leqslant \bar{\gamma}+\epsilon_i,\ \forall i\in[m],$$

其中 $\bar{\gamma}$ 是间隔均值, η 和 λ 是权衡各项重要性的超参数.

由于预测结果只与决策函数值的符号有关, 因此可线性拉伸 $\boldsymbol{w}$ 使得间隔均值固定为 1; 间隔大于 1 和小于 1 虽然都偏离了均值, 但后者比前者更容易出现预测错误, 因此这两种偏差不是对等的; 由于只有间隔正好等于 1 才没有损失, 因此最终分类超平面几乎由全部样本构成. 为了控制模型稀疏度, ODM 采用了 θ-不敏感损失 [19], 最终形式化为

$$\begin{aligned}\min_{\boldsymbol{w},\xi_i,\epsilon_i} \ & \frac{1}{2}\|\boldsymbol{w}\|^2 + \frac{\lambda}{2m}\sum_{i\in[m]}(\xi_i^2+\nu\epsilon_i^2),\\ \text{s.t. } & y_i\langle\boldsymbol{w},\phi(\boldsymbol{x}_i)\rangle \geqslant 1-\theta-\xi_i,\ y_i\langle\boldsymbol{w},\phi(\boldsymbol{x}_i)\rangle \leqslant 1+\theta+\epsilon_i,\ \forall i\in[m],\end{aligned} \tag{1}$$

其中 ν 是权衡两种不同偏差的超参数, θ 是控制不敏感损失带宽度的超参数.

2.2 基于间隔的聚类

基于间隔的聚类通过联合优化簇标记赋值和目标间隔物理量实现聚类, 由于簇标记都是整型变量, 因此最终形式化得到的是混合整数规划. 这类问题理论上是 NP-hard 的, 直接求解不现实, 于是很多近似求解的方法被提了出来, 大致可以分为两类.

(1) 放松成凸优化问题再求解. 由于簇标记 $\boldsymbol{y}$ 总是以 $\boldsymbol{y}\boldsymbol{y}^{\mathrm{T}}$ 的外积形式出现, Xu 等 [6] 直接将其写成一个半正定矩阵, 从而得到一个半定规划 (semi-definite programming, SDP). 在此基础上, Valizadegan 等 [7] 提出了一个新的放松, 虽然还是基于 SDP, 但变量个数从 $\mathcal{O}(m^2)$ 大幅减少到 $\mathcal{O}(m)$. Li 等 [8] 提出一个基于极大极小不等式 [20] 的放松, 并在理论上证明了它比基于 SDP 的放松更紧, 最终求解该问题可转化为一个多核学习问题 [21]. 以上这些算法优化的间隔物理量均是最小间隔. 近期, Zhang 等 [11] 提出以间隔分布为优化目标的聚类算法 ODMC (optimal margin distribution machine for clustering), 该方法基于最新的间隔理论, 因此取得了比之前方法更好的性能.

(2) 直接用非凸优化算法求局部最小. Zhang 等 [9, 22] 采用交替优化的算法, 即每轮先固定簇标记 $\boldsymbol{y}$, 求解分类超平面; 再根据分类超平面, 重新优化簇标记赋值, 如此交替迭代直至收敛. Zhao 等 [10] 先将簇标记 $\boldsymbol{y}$ 用绝对值函数吸收掉, 虽然整型优化变量没有了, 但这将使得约束变得非凸, 接着采用 CCCP (concave-convex procedure) 的方法 [23], 将非凸约束写成两个凸函数的差, 并将后者进一步用一阶 Taylor 展式代替, 从而将整体变成一个较易求解的问题. 该类方法只能求得一个局部最小, 因此需多次初始化取最优的模型.

还有很多研究者尝试对基于间隔的聚类方法做进一步推广, 例如 Zhou 等 [24] 考虑了数据中存在隐变量的情况. Niu 等 [25] 提出一个类似的准则叫基于容量的聚类. Vijaya Saradhi 等 [26] 引入了增量学习的设定.

2.3 半监督聚类

现实聚类任务中, 样本间的成对约束远比样本的类别标记更容易获得, 且根据类别标记可以导出成对约束, 因此本文只考虑成对约束形式的监督信息. 根据对其利用方式的不同, 现有的 SSC 算法可分为两类.

(1) 使用约束来指导最优聚类结构的搜索. Wagstaff 等[27] 提出 COPKMEANS, 其修改了 k 均值算法每轮迭代时的簇标记赋值操作, 除寻找最近的簇中心, 还需满足所有约束. 为了克服 COPKMEANS 的硬约束限制, Davidson 等[28] 提出 CVQE (constrained vector quantization error), 为每个约束设计了基于样本距离的损失, 并将其加入到 VQE 的目标函数中进行优化, 从而 "软化" 了约束. CVQE 每轮迭代需为每个约束测试所有簇标记对, 因此时间复杂度为 $\mathcal{O}(k^2)$. 为减少计算开销, Pelleg 等[29] 提出 LCVQE (linear-time CVQE), 通过修改约束对应的距离损失使得每轮只需检测 3 个簇标记对. Basu 等[30] 提出 PCKMEANS, 直接将约束满足与否作为 0-1 损失加入到目标函数中, 同时为每个约束设置了权重, 作为超参数由用户输入. 除了将约束集成进 k 均值和 VQE 算法, 也有工作[31,32] 考虑了高斯 (Gauss) 混合模型, 该类算法可通过 EM 算法求解.

(2) 根据约束先进行度量学习 (metric learning), 再用常规的聚类算法进行聚类. Xing 等[33] 将度量学习形式化成一个最小化问题, 目标函数是所有 "必连" 约束的两个样本的距离之和, 约束是所有 "勿连" 约束的两个样本的距离之和不小于一个给定常数. Hertz 等[34] 将给定约束对应的样本对作为样本, 约束类型作为类别标记, 用以高斯混合模型为基分类器的 Adaboost 算法[35] 学习两者之间的映射关系, 从而对任意两个样本都可预测其约束关系. Liu 等[36] 引入核映射以期在新的特征空间中 "必连" 约束的两个样本距离尽可能近, 同时 "勿连" 约束的两个样本距离尽可能远, 最终将问题转化为对核矩阵的学习.

还有很多研究者引入了更多的假设, 例如 Yi 等[37] 假设成对约束是以流的形式得到的, 此时需进行增量式的半监督聚类. Cohn 等[38] 假设有一个可以交互的领域专家能给出任意两个样本间的约束类型, 从而用主动学习[39] 的机制进行半监督聚类. 近年来, 随着深度学习[40] 的兴起, 亦有研究者将其引入到了半监督聚类, 其基本想法是利用自动编码器学习一个好的特征表示[41], 再利用传统的半监督算法进行聚类, 只需将编码器的重构损失和破坏约束的损失合起来, 就能实现端到端的优化.

3 ODMSSC

ODMSSC 采用第 2.3 小节中的第 1 种策略, 将监督信息引入簇标记赋值和间隔分布的联合优化, 加速对最优聚类结构的搜索. 首先将式 (1) 中的 y_i 作为优化变量可得

$$\begin{aligned}&\min_{\hat{\boldsymbol{y}}\in\mathcal{B}}\min_{\boldsymbol{w},\xi_i,\epsilon_i}\ \frac{1}{2}\|\boldsymbol{w}\|^2+\frac{\lambda}{2m}\sum_{i\in[m]}(\xi_i^2+\nu\epsilon_i^2),\\&\text{s.t. }\hat{y}_i\langle\boldsymbol{w},\phi(\boldsymbol{x}_i)\rangle\geqslant 1-\theta-\xi_i,\ \hat{y}_i\langle\boldsymbol{w},\phi(\boldsymbol{x}_i)\rangle\leqslant 1+\theta+\epsilon_i,\ \forall i\in[m],\end{aligned}\tag{2}$$

其中 $\mathcal{B}=\{\hat{\boldsymbol{y}}=[\hat{y}_1;\ldots;\hat{y}_m]\in\{1,-1\}^m\mid \hat{y}_i=\hat{y}_j,\ (i,j)\in\mathcal{M},\ \hat{y}_k\neq\hat{y}_l,\ (k,l)\in\mathcal{C}\}$ 是候选簇标记赋值集合, $\mathcal{M}$ 是 "必连" 约束集合, $\mathcal{C}$ 是 "勿连" 约束集合.

引入拉格朗日乘子变量 $\boldsymbol{\alpha}=[\alpha_1;\ldots;\alpha_m]$, $\boldsymbol{\beta}=[\beta_1;\ldots;\beta_m]$, 根据 KKT (Karush-Kuhn-Tucker) 条件

$$\boldsymbol{w}=\sum_{i\in[m]}(\alpha_i-\beta_i)\hat{y}_i\phi(\boldsymbol{x}_i),\quad \lambda\xi_i=m\alpha_i,\quad \lambda\nu\epsilon_i=m\beta_i,$$

并记 $\boldsymbol{\delta}=\boldsymbol{\alpha}-\boldsymbol{\beta}$, 可知式 (2) 的对偶问题为

$$\min_{\hat{\boldsymbol{y}}\in\mathcal{B}}\max_{\boldsymbol{\alpha},\boldsymbol{\beta}\succeq\boldsymbol{0}}-\frac{1}{2}\boldsymbol{\delta}^{\mathrm{T}}(\boldsymbol{K}\odot\hat{\boldsymbol{y}}\hat{\boldsymbol{y}}^{\mathrm{T}})\boldsymbol{\delta}-\frac{m}{2\lambda}\begin{bmatrix}\boldsymbol{\alpha}\\\boldsymbol{\beta}\end{bmatrix}^{\mathrm{T}}\begin{bmatrix}\boldsymbol{I}&\\&\boldsymbol{I}/\nu\end{bmatrix}\begin{bmatrix}\boldsymbol{\alpha}\\\boldsymbol{\beta}\end{bmatrix}-\begin{bmatrix}(\theta-1)\boldsymbol{e}\\(\theta+1)\boldsymbol{e}\end{bmatrix}^{\mathrm{T}}\begin{bmatrix}\boldsymbol{\alpha}\\\boldsymbol{\beta}\end{bmatrix},\tag{3}$$

其中 $\boldsymbol{K}$ 是核矩阵满足 $[\boldsymbol{K}]_{ij}=\langle\phi(\boldsymbol{x}_i),\phi(\boldsymbol{x}_j)\rangle$, $\boldsymbol{I}$ 是 m 阶单位阵, $\odot$ 表示元素按位乘, $\boldsymbol{e}$ 是全 1 向量. 由于外层问题的优化变量 $\hat{\boldsymbol{y}}$ 是离散的, 式 (3) 是一个混合整数规划, 难以直接求解. 现交换 $\min_{\hat{\boldsymbol{y}}\in\mathcal{B}}$ 和 $\max_{\boldsymbol{\alpha},\boldsymbol{\beta}\succeq\boldsymbol{0}}$ 的顺序, 即优化式 (3) 的下界, 并注意

$$\min_{\hat{\boldsymbol{y}}\in\mathcal{B}} f(\hat{\boldsymbol{y}})=\min\left\{f(\hat{\boldsymbol{y}}_1),\ldots,f(\hat{\boldsymbol{y}}_{|\mathcal{B}|})\right\}=\min_{\boldsymbol{\mu}\in\Delta^{|\mathcal{B}|}}\sum_{t:\hat{\boldsymbol{y}}_t\in\mathcal{B}}\mu_t f(\hat{\boldsymbol{y}}_t),$$

其中 $\Delta^{|\mathcal{B}|}=\{\boldsymbol{\mu}=[\mu_1;\ldots;\mu_{|\mathcal{B}|}]\mid\boldsymbol{e}^{\mathrm{T}}\boldsymbol{\mu}=1,\ \boldsymbol{\mu}\succeq\boldsymbol{0}\}$ 是 $|\mathcal{B}|$ 维单纯形, 于是可得连续优化问题:

$$\max_{\boldsymbol{\alpha},\boldsymbol{\beta}\succeq\boldsymbol{0}}\min_{\boldsymbol{\mu}\in\Delta^{|\mathcal{B}|}}-\frac{1}{2}\boldsymbol{\delta}^{\mathrm{T}}\left(\sum_{t:\hat{\boldsymbol{y}}_t\in\mathcal{B}}\mu_t\boldsymbol{K}\odot\hat{\boldsymbol{y}}_t\hat{\boldsymbol{y}}_t^{\mathrm{T}}\right)\boldsymbol{\delta}-\frac{m}{2\lambda}\begin{bmatrix}\boldsymbol{\alpha}\\\boldsymbol{\beta}\end{bmatrix}^{\mathrm{T}}\begin{bmatrix}\boldsymbol{I}&\\&\boldsymbol{I}/\nu\end{bmatrix}\begin{bmatrix}\boldsymbol{\alpha}\\\boldsymbol{\beta}\end{bmatrix}-\begin{bmatrix}(\theta-1)\boldsymbol{e}\\(\theta+1)\boldsymbol{e}\end{bmatrix}^{\mathrm{T}}\begin{bmatrix}\boldsymbol{\alpha}\\\boldsymbol{\beta}\end{bmatrix}.$$

注意 $\boldsymbol{\delta}=\boldsymbol{\alpha}-\boldsymbol{\beta}=[\boldsymbol{I};-\boldsymbol{I}]^{\mathrm{T}}[\boldsymbol{\alpha};\boldsymbol{\beta}]$ 是关于 $[\boldsymbol{\alpha};\boldsymbol{\beta}]$ 的线性变换, 易知上式的目标函数是关于 $[\boldsymbol{\alpha};\boldsymbol{\beta}]$ 的凹函数、关于 $\boldsymbol{\mu}$ 的凸函数, 于是由鞍点定理 [42] 知其等价于

$$\min_{\boldsymbol{\mu}\in\Delta^{|\mathcal{B}|}}\max_{\boldsymbol{\alpha},\boldsymbol{\beta}\succeq\boldsymbol{0}}-\frac{1}{2}\boldsymbol{\delta}^{\mathrm{T}}\left(\sum_{t:\hat{\boldsymbol{y}}_t\in\mathcal{B}}\mu_t\boldsymbol{K}\odot\hat{\boldsymbol{y}}_t\hat{\boldsymbol{y}}_t^{\mathrm{T}}\right)\boldsymbol{\delta}-\frac{m}{2\lambda}\begin{bmatrix}\boldsymbol{\alpha}\\\boldsymbol{\beta}\end{bmatrix}^{\mathrm{T}}\begin{bmatrix}\boldsymbol{I}&\\&\boldsymbol{I}/\nu\end{bmatrix}\begin{bmatrix}\boldsymbol{\alpha}\\\boldsymbol{\beta}\end{bmatrix}-\begin{bmatrix}(\theta-1)\boldsymbol{e}\\(\theta+1)\boldsymbol{e}\end{bmatrix}^{\mathrm{T}}\begin{bmatrix}\boldsymbol{\alpha}\\\boldsymbol{\beta}\end{bmatrix},\tag{4}$$

这就是 ODMSSC 的对偶问题.

下面推导式 (4) 对应的原问题, 注意

$$\left[\sum_{t:\hat{\boldsymbol{y}}_t\in\mathcal{B}}\mu_t\boldsymbol{K}\odot\hat{\boldsymbol{y}}_t\hat{\boldsymbol{y}}_t^{\mathrm{T}}\right]_{ij}=\sum_{t:\hat{\boldsymbol{y}}_t\in\mathcal{B}}\mu_t[\boldsymbol{K}]_{ij}[\hat{\boldsymbol{y}}_t]_i[\hat{\boldsymbol{y}}_t]_j=\left\langle\begin{bmatrix}\sqrt{\mu_1}[\hat{\boldsymbol{y}}_1]_i\phi(\boldsymbol{x}_i)\\\vdots\\\sqrt{\mu_{|\mathcal{B}|}}[\hat{\boldsymbol{y}}_{|\mathcal{B}|}]_i\phi(\boldsymbol{x}_i)\end{bmatrix},\begin{bmatrix}\sqrt{\mu_1}[\hat{\boldsymbol{y}}_1]_j\phi(\boldsymbol{x}_j)\\\vdots\\\sqrt{\mu_{|\mathcal{B}|}}[\hat{\boldsymbol{y}}_{|\mathcal{B}|}]_j\phi(\boldsymbol{x}_j)\end{bmatrix}\right\rangle,$$

引入辅助样本 $\tilde{\boldsymbol{x}}_i=[\sqrt{\mu_1}[\hat{\boldsymbol{y}}_1]_i\phi(\boldsymbol{x}_i);\ldots;\sqrt{\mu_{|\mathcal{B}|}}[\hat{\boldsymbol{y}}_{|\mathcal{B}|}]_i\phi(\boldsymbol{x}_i)]$ 和辅助核矩阵 $\tilde{\boldsymbol{K}}$ 满足 $[\tilde{\boldsymbol{K}}]_{ij}=\langle\tilde{\boldsymbol{x}}_i,\tilde{\boldsymbol{x}}_j\rangle$, 则式 (4) 可写成

$$\min_{\boldsymbol{\mu}\in\Delta^{|\mathcal{B}|}}\max_{\boldsymbol{\alpha},\boldsymbol{\beta}\succeq\boldsymbol{0}}-\frac{1}{2}\boldsymbol{\delta}^{\mathrm{T}}\tilde{\boldsymbol{K}}\boldsymbol{\delta}-\frac{m}{2\lambda}\begin{bmatrix}\boldsymbol{\alpha}\\\boldsymbol{\beta}\end{bmatrix}^{\mathrm{T}}\begin{bmatrix}\boldsymbol{I}&\\&\boldsymbol{I}/\nu\end{bmatrix}\begin{bmatrix}\boldsymbol{\alpha}\\\boldsymbol{\beta}\end{bmatrix}-\begin{bmatrix}(\theta-1)\boldsymbol{e}\\(\theta+1)\boldsymbol{e}\end{bmatrix}^{\mathrm{T}}\begin{bmatrix}\boldsymbol{\alpha}\\\boldsymbol{\beta}\end{bmatrix},\tag{5}$$

注意 $\tilde{\boldsymbol{K}}=\tilde{\boldsymbol{K}}\odot\boldsymbol{e}\boldsymbol{e}^{\mathrm{T}}$, 对比式 (2) 和 (3) 的形式不难看出式 (5) 对应的原问题是

$$\begin{aligned}&\min_{\boldsymbol{\mu}\in\Delta^{|\mathcal{B}|}}\min_{\tilde{\boldsymbol{w}},\xi_i,\epsilon_i}\ \frac{1}{2}\|\tilde{\boldsymbol{w}}\|^2+\frac{\lambda}{2m}\sum_{i\in[m]}(\xi_i^2+\nu\epsilon_i^2),\\&\text{s.t. }\langle\tilde{\boldsymbol{w}},\tilde{\boldsymbol{x}}_i\rangle\geqslant1-\theta-\xi_i,\ \langle\tilde{\boldsymbol{w}},\tilde{\boldsymbol{x}}_i\rangle\leqslant1+\theta+\epsilon_i,\ \forall i\in[m].\end{aligned}\tag{6}$$

且有 KKT 条件 $\tilde{\boldsymbol{w}}=\sum_{i\in[m]}(\alpha_i-\beta_i)\tilde{\boldsymbol{x}}_i$, $\lambda\xi_i=m\alpha_i$, $\lambda\nu\epsilon_i=m\beta_i$. 由于外层优化变量 $\boldsymbol{\mu}$ 隐藏在 $\tilde{\boldsymbol{x}}_i$ 中, 现将其显式地写出来, 设 $\tilde{\boldsymbol{w}}=[\boldsymbol{w}_1/\sqrt{\mu_1};\ldots;\boldsymbol{w}_{|\mathcal{B}|}/\sqrt{\mu_{|\mathcal{B}|}}]$, 则 $\langle\tilde{\boldsymbol{w}},\tilde{\boldsymbol{x}}_i\rangle=\sum_{t:\hat{\boldsymbol{y}}_t\in\mathcal{B}}[\hat{\boldsymbol{y}}_t]_i\langle\boldsymbol{w}_t,\phi(\boldsymbol{x}_i)\rangle$, 于是式 (6) 可进一步写成

$$\begin{aligned}&\min_{\boldsymbol{\mu}\in\Delta^{|\mathcal{B}|}}\min_{\boldsymbol{w}_t,\xi_i,\epsilon_i}\ \frac{1}{2}\sum_{t:\hat{\boldsymbol{y}}_t\in\mathcal{B}}\frac{\|\boldsymbol{w}_t\|^2}{\mu_t}+\frac{\lambda}{2m}\sum_{i\in[m]}(\xi_i^2+\nu\epsilon_i^2),\\&\text{s.t. }\sum_{t:\hat{\boldsymbol{y}}_t\in\mathcal{B}}[\hat{\boldsymbol{y}}_t]_i\langle\boldsymbol{w}_t,\phi(\boldsymbol{x}_i)\rangle\geqslant1-\theta-\xi_i,\ \sum_{t:\hat{\boldsymbol{y}}_t\in\mathcal{B}}[\hat{\boldsymbol{y}}_t]_i\langle\boldsymbol{w}_t,\phi(\boldsymbol{x}_i)\rangle\leqslant1+\theta+\epsilon_i,\ \forall i\in[m],\end{aligned}\tag{7}$$

这就是 ODMSSC 的原问题.

4 优化

我们采用交替优化进行求解, 即每轮先固定 $\boldsymbol{\mu}$, 然后优化 $\boldsymbol{\alpha}$ 和 $\boldsymbol{\beta}$; 再固定 $\boldsymbol{\alpha}$ 和 $\boldsymbol{\beta}$, 然后优化 $\boldsymbol{\mu}$.

4.1 交替优化

当 $\boldsymbol{\mu}$ 固定时, $\tilde{\boldsymbol{K}}$ 亦固定, 根据式 (5), $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ 优化子问题为

$$\max_{\boldsymbol{\alpha},\boldsymbol{\beta}\succeq\mathbf{0}} -\frac{1}{2}\begin{bmatrix}\boldsymbol{\alpha}\\ \boldsymbol{\beta}\end{bmatrix}^{\mathrm{T}}\begin{bmatrix}\boldsymbol{I}\\ -\boldsymbol{I}\end{bmatrix}\tilde{\boldsymbol{K}}\begin{bmatrix}\boldsymbol{I}\\ -\boldsymbol{I}\end{bmatrix}^{\mathrm{T}}\begin{bmatrix}\boldsymbol{\alpha}\\ \boldsymbol{\beta}\end{bmatrix}-\frac{m}{2\lambda}\begin{bmatrix}\boldsymbol{\alpha}\\ \boldsymbol{\beta}\end{bmatrix}^{\mathrm{T}}\begin{bmatrix}\boldsymbol{I} & \\ & \boldsymbol{I}/\nu\end{bmatrix}\begin{bmatrix}\boldsymbol{\alpha}\\ \boldsymbol{\beta}\end{bmatrix}-\begin{bmatrix}(\theta-1)\boldsymbol{e}\\ (\theta+1)\boldsymbol{e}\end{bmatrix}^{\mathrm{T}}\begin{bmatrix}\boldsymbol{\alpha}\\ \boldsymbol{\beta}\end{bmatrix}.$$

这是非负象限上的凸二次规划问题, 采用投影梯度法即可求解, 此时有闭式解:

$$\begin{bmatrix}\boldsymbol{\alpha}\\ \boldsymbol{\beta}\end{bmatrix}\leftarrow\max\left\{\mathbf{0},\begin{bmatrix}\boldsymbol{\alpha}\\ \boldsymbol{\beta}\end{bmatrix}-\eta\left(\begin{bmatrix}\tilde{\boldsymbol{K}} & -\tilde{\boldsymbol{K}}\\ -\tilde{\boldsymbol{K}} & \tilde{\boldsymbol{K}}\end{bmatrix}\begin{bmatrix}\boldsymbol{\alpha}\\ \boldsymbol{\beta}\end{bmatrix}+\frac{m}{\lambda}\begin{bmatrix}\boldsymbol{I} & \\ & \boldsymbol{I}/\nu\end{bmatrix}\begin{bmatrix}\boldsymbol{\alpha}\\ \boldsymbol{\beta}\end{bmatrix}+\begin{bmatrix}(\theta-1)\boldsymbol{e}\\ (\theta+1)\boldsymbol{e}\end{bmatrix}\right)\right\},\tag{8}$$

其中 η 是学习率. 求得 $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ 后, 由 $\lambda\xi_i=m\alpha_i$, $\lambda\nu\epsilon_i=m\beta_i$ 可得 ϵ_i, ξ_i. 又由

$$\begin{bmatrix}\boldsymbol{w}_1/\sqrt{\mu_1}\\ \vdots\\ \boldsymbol{w}_{|\mathcal{B}|}/\sqrt{\mu_{|\mathcal{B}|}}\end{bmatrix}=\tilde{\boldsymbol{w}}=\sum_{i\in[m]}(\alpha_i-\beta_i)\tilde{\boldsymbol{x}}_i=\sum_{i\in[m]}(\alpha_i-\beta_i)\begin{bmatrix}\sqrt{\mu_1}[\hat{\boldsymbol{y}}_1]_i\phi(\boldsymbol{x}_i)\\ \vdots\\ \sqrt{\mu_{|\mathcal{B}|}}[\hat{\boldsymbol{y}}_{|\mathcal{B}|}]_i\phi(\boldsymbol{x}_i)\end{bmatrix}\tag{9}$$

可得 $\boldsymbol{w}_t=\sum_{i\in[m]}(\alpha_i-\beta_i)\mu_t[\hat{\boldsymbol{y}}_t]_i\phi(\boldsymbol{x}_i)$.

当 $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ 固定时, $\boldsymbol{w}_t$, ϵ_i, ξ_i 亦固定, 根据式 (7), $\boldsymbol{\mu}$ 优化子问题为

$$\min_{\boldsymbol{\mu}\in\Delta^{|\mathcal{B}|}}\sum_{t:\hat{\boldsymbol{y}}_t\in\mathcal{B}}\frac{\|\boldsymbol{w}_t\|^2}{\mu_t}.$$

注意 $\sum_{t:\hat{\boldsymbol{y}}_t\in\mathcal{B}}\mu_t=1$, 由柯西 – 施瓦茨 (Cauchy-Schwarz) 不等式知

$$\sum_{t:\hat{\boldsymbol{y}}_t\in\mathcal{B}}\frac{\|\boldsymbol{w}_t\|^2}{\mu_t}=\left(\sum_{t:\hat{\boldsymbol{y}}_t\in\mathcal{B}}\frac{\|\boldsymbol{w}_t\|^2}{\mu_t}\right)\left(\sum_{t:\hat{\boldsymbol{y}}_t\in\mathcal{B}}\mu_t\right)\geqslant\left(\sum_{t:\hat{\boldsymbol{y}}_t\in\mathcal{B}}\|\boldsymbol{w}_t\|\right)^2.$$

取等号的条件是 $\|\boldsymbol{w}_t\|/\mu_t$ 为常数, 不妨设为 k, 于是

$$\mu_t=\frac{\|\boldsymbol{w}_t\|}{k}=\frac{\|\boldsymbol{w}_t\|}{k\sum_{i:\hat{\boldsymbol{y}}_i\in\mathcal{B}}\mu_i}=\frac{\|\boldsymbol{w}_t\|}{\sum_{i:\hat{\boldsymbol{y}}_i\in\mathcal{B}}\|\boldsymbol{w}_i\|},\tag{10}$$

即 $\boldsymbol{\mu}$ 优化子问题有闭式解.

4.2 算法实现

本小节给出算法的具体实现, 注意 $\boldsymbol{K}\odot\hat{\boldsymbol{y}}\hat{\boldsymbol{y}}^{\mathrm{T}}=\mathrm{diag}(\hat{\boldsymbol{y}})\boldsymbol{K}\mathrm{diag}(\hat{\boldsymbol{y}})$ 以及 $\mathrm{diag}(\hat{\boldsymbol{y}})\boldsymbol{\delta}=\mathrm{diag}(\boldsymbol{\delta})\hat{\boldsymbol{y}}$, 故式 (3) 关于 $\hat{\boldsymbol{y}}$ 的优化问题可重写为

$$\max_{\hat{\boldsymbol{y}}\in\mathcal{B}}\boldsymbol{\delta}^{\mathrm{T}}\mathrm{diag}(\hat{\boldsymbol{y}})\boldsymbol{K}\mathrm{diag}(\hat{\boldsymbol{y}})\boldsymbol{\delta}=\max_{\hat{\boldsymbol{y}}\in\mathcal{B}}\hat{\boldsymbol{y}}^{\mathrm{T}}\mathrm{diag}(\boldsymbol{\delta})\boldsymbol{K}\mathrm{diag}(\boldsymbol{\delta})\hat{\boldsymbol{y}}=\max_{\hat{\boldsymbol{y}}\in\mathcal{B}}\hat{\boldsymbol{y}}^{\mathrm{T}}\boldsymbol{H}\hat{\boldsymbol{y}},$$

其中 $\boldsymbol{H}=\mathrm{diag}(\boldsymbol{\delta})\boldsymbol{K}\mathrm{diag}(\boldsymbol{\delta})$ 是半正定矩阵. 由于 $\hat{\boldsymbol{y}}$ 是整型变量, 难以直接求解, 我们引入高维向量 $\boldsymbol{\mu}\in\Delta^{|\mathcal{B}|}$ 将问题连续化, 其每一维对应一个候选的簇标记赋值, 因此维度与样本数呈指数关系. 实际处理时无需维护这么高维的向量, 一个稀疏的 $\boldsymbol{\mu}$ 就足以以很高的精度接近最优解. 故算法总体结构可设计为内外双层循环, 外层循环不断扩充 $\boldsymbol{\mu}$ 的非零维, 内层循环在固定 $\boldsymbol{\mu}$ 的非零维的情况下, 做第 4.1 小节的交替优化, 并输出最优目标函数值, 当相邻两轮外层循环得到的最优目标函数值无显著变化时, 停止对 $\boldsymbol{\mu}$ 的非零维的扩充, 结束整个优化过程.

设第 T 轮外层循环 $\boldsymbol{\mu}$ 的非零维下标集合为 $\mathcal{I}_T$, 候选簇标记赋值集合 $\mathcal{A}_T=\{\hat{\boldsymbol{y}}_t\in\mathcal{B}\mid t\in\mathcal{I}_T\}$, 由于 $\mathcal{I}_1\subseteq\mathcal{I}_2\subseteq\cdots\subseteq[|\mathcal{B}|]$, 因此 $\mathcal{A}_1\subseteq\mathcal{A}_2\subseteq\cdots\subseteq\mathcal{B}$, 故

$$\max_{\hat{\boldsymbol{y}}\in\mathcal{A}_1}\hat{\boldsymbol{y}}^{\mathrm{T}}\boldsymbol{H}\hat{\boldsymbol{y}}<\max_{\hat{\boldsymbol{y}}\in\mathcal{A}_2}\hat{\boldsymbol{y}}^{\mathrm{T}}\boldsymbol{H}\hat{\boldsymbol{y}}<\cdots<\max_{\hat{\boldsymbol{y}}\in\mathcal{B}}\hat{\boldsymbol{y}}^{\mathrm{T}}\boldsymbol{H}\hat{\boldsymbol{y}},\tag{11}$$

即 $\{\max_{\hat{\boldsymbol{y}}\in\mathcal{A}_t}\hat{\boldsymbol{y}}^{\mathrm{T}}\boldsymbol{H}\hat{\boldsymbol{y}}\}_{t=1,2,\ldots}$ 是单调递增序列, 又 $\max_{\hat{\boldsymbol{y}}\in\mathcal{B}}\hat{\boldsymbol{y}}^{\mathrm{T}}\boldsymbol{H}\hat{\boldsymbol{y}}$ 显然为其上界, 根据单调收敛定理可知该序列有极限, 算法必收敛. 换言之, 扩充 $\boldsymbol{\mu}$ 的非零维的过程相当于构造式 (11) 的问题序列逐渐逼近原问题的过程. 欲使该序列尽快收敛, 扩充 $\mathcal{A}_T$ 时应选择尽可能使得目标函数值增大的 $\hat{\boldsymbol{y}}$, 但这是一个凸函数求极大值的问题, 难以直接求解, 下面给出一个简单且可使得式 (11) 的序列严格单调增的方法. 记

$$\bar{\boldsymbol{y}}=\operatorname*{argmax}_{\hat{\boldsymbol{y}}\in\mathcal{A}_T}\hat{\boldsymbol{y}}^{\mathrm{T}}\boldsymbol{H}\hat{\boldsymbol{y}},\quad \tilde{\boldsymbol{y}}=\operatorname*{argmax}_{\hat{\boldsymbol{y}}\in\mathcal{B}}\hat{\boldsymbol{y}}^{\mathrm{T}}\boldsymbol{H}\bar{\boldsymbol{y}},\tag{12}$$

于是只要 $\tilde{\boldsymbol{y}}^{\mathrm{T}}\boldsymbol{H}\bar{\boldsymbol{y}}\neq\bar{\boldsymbol{y}}^{\mathrm{T}}\boldsymbol{H}\bar{\boldsymbol{y}}$, 就有 $\tilde{\boldsymbol{y}}^{\mathrm{T}}\boldsymbol{H}\tilde{\boldsymbol{y}}>\bar{\boldsymbol{y}}^{\mathrm{T}}\boldsymbol{H}\bar{\boldsymbol{y}}$. 用反证法, 设 $\tilde{\boldsymbol{y}}^{\mathrm{T}}\boldsymbol{H}\tilde{\boldsymbol{y}}\leqslant\bar{\boldsymbol{y}}^{\mathrm{T}}\boldsymbol{H}\bar{\boldsymbol{y}}$, 则有

$$0\leqslant(\tilde{\boldsymbol{y}}-\bar{\boldsymbol{y}})^{\mathrm{T}}\boldsymbol{H}(\tilde{\boldsymbol{y}}-\bar{\boldsymbol{y}})=\tilde{\boldsymbol{y}}^{\mathrm{T}}\boldsymbol{H}\tilde{\boldsymbol{y}}+\bar{\boldsymbol{y}}^{\mathrm{T}}\boldsymbol{H}\bar{\boldsymbol{y}}-2\tilde{\boldsymbol{y}}^{\mathrm{T}}\boldsymbol{H}\bar{\boldsymbol{y}}\leqslant2(\bar{\boldsymbol{y}}^{\mathrm{T}}\boldsymbol{H}\bar{\boldsymbol{y}}-\tilde{\boldsymbol{y}}^{\mathrm{T}}\boldsymbol{H}\bar{\boldsymbol{y}})<0$$

矛盾, 故按照式 (12) 得到的 $\tilde{\boldsymbol{y}}$ 可使得式 (11) 的序列严格单调增.

注意 $\mathcal{A}_T$ 中只有有限个元素, 因此 $\bar{\boldsymbol{y}}$ 不难求解, 遍历全部元素取令目标函数最大的 $\hat{\boldsymbol{y}}$ 即可. 求解 $\tilde{\boldsymbol{y}}$ 是整数线性规划, 由于 $\hat{\boldsymbol{y}}\in\{1,-1\}^m$, 故若不考虑约束关系, 则有闭式解 $\tilde{\boldsymbol{y}}=\mathrm{sgn}(\boldsymbol{H}\bar{\boldsymbol{y}})$. 当 $\tilde{\boldsymbol{y}}$ 不满足某个约束时, 不妨设为 (i,j), 此时需改变 $[\tilde{\boldsymbol{y}}]_i$ 和 $[\tilde{\boldsymbol{y}}]_j$ 之一的符号, 由于该操作可能会破坏已经满足的约束, 因此设要改变的是 $[\tilde{\boldsymbol{y}}]_i$, 那么还需同时改变与 $\boldsymbol{x}_i$ 有 "必连" 约束和 "勿连" 约束关系的所有样本的簇标记赋值. 具体来说, 记 $\bar{\mathcal{M}}=\mathcal{M}\vee\{(p,q)\mid(q,p)\in\mathcal{M}\}$ 为 $\mathcal{M}$ 的对称闭包, 则 $\bar{\mathcal{M}}$ 的传递闭包

$$\tilde{\mathcal{M}}=\bar{\mathcal{M}}\vee\{(p,q)\mid\exists\{z_i\}_{i\in[n]}\to(p,z_1)\in\bar{\mathcal{M}}\wedge(z_1,z_2)\in\bar{\mathcal{M}}\wedge\cdots\wedge(z_n,q)\in\bar{\mathcal{M}}\},\tag{13}$$

包含了由 $\mathcal{M}$ 诱导出的全部 "必连" 约束. 记

$$\mathcal{M}_i=\{j\mid(i,j)\in\tilde{\mathcal{M}}\vee(j,i)\in\tilde{\mathcal{M}}\},\quad \mathcal{C}_i=\{j\mid((i,k)\in\mathcal{C}\vee(k,i)\in\mathcal{C})\wedge j\in\mathcal{M}_k\},\tag{14}$$

分别为与 $\boldsymbol{x}_i$ 构成 "必连" 约束关系和 "勿连" 约束关系的全部样本的下标构成的集合, 记

$$\tilde{\boldsymbol{y}}_i=\tilde{\boldsymbol{y}}-2\cdot\mathrm{sgn}([\tilde{\boldsymbol{y}}]_i)(\boldsymbol{e}_{\{i\}}+\boldsymbol{e}_{\mathcal{M}_i}-\boldsymbol{e}_{\mathcal{C}_i}),\quad \tilde{\boldsymbol{y}}_j=\tilde{\boldsymbol{y}}-2\cdot\mathrm{sgn}([\tilde{\boldsymbol{y}}]_j)(\boldsymbol{e}_{\{j\}}+\boldsymbol{e}_{\mathcal{M}_j}-\boldsymbol{e}_{\mathcal{C}_j}),\tag{15}$$

其中 $\boldsymbol{e}_{\mathcal{A}}$ 是 m 维 0-1 向量满足 $[\boldsymbol{e}_{\mathcal{A}}]_k=1_{k\in\mathcal{A}}$, 则 $\tilde{\boldsymbol{y}}_i$, $\tilde{\boldsymbol{y}}_j$ 分别为在满足所有约束的情况下改变 $[\tilde{\boldsymbol{y}}]_i$, $[\tilde{\boldsymbol{y}}]_j$ 得到的簇标记赋值, 剩下只需比较 $\tilde{\boldsymbol{y}}_i^{\mathrm{T}}\boldsymbol{H}\tilde{\boldsymbol{y}}_i$ 和 $\tilde{\boldsymbol{y}}_j^{\mathrm{T}}\boldsymbol{H}\tilde{\boldsymbol{y}}_j$ 的大小, 并令

$$\tilde{\boldsymbol{y}}=\begin{cases}\tilde{\boldsymbol{y}}_i, & \text{if } \tilde{\boldsymbol{y}}_i^{\mathrm{T}}\boldsymbol{H}\tilde{\boldsymbol{y}}_i\geqslant\tilde{\boldsymbol{y}}_j^{\mathrm{T}}\boldsymbol{H}\tilde{\boldsymbol{y}}_j,\\ \tilde{\boldsymbol{y}}_j, & \text{if } \tilde{\boldsymbol{y}}_i^{\mathrm{T}}\boldsymbol{H}\tilde{\boldsymbol{y}}_i<\tilde{\boldsymbol{y}}_j^{\mathrm{T}}\boldsymbol{H}\tilde{\boldsymbol{y}}_j.\end{cases}\tag{16}$$

综上, ODMSSC 算法的具体执行步骤如算法 1 所示.

Algorithm 1 ODMSSC

Input: Data set $\mathcal{S}=\{\boldsymbol{x}_i\}_{i\in[m]}$, "must-link" constraint set $\mathcal{M}$, "cannot-link" constraint set $\mathcal{C}$, maximum iteration number T, stopping criteria τ.

Output: $\tilde{\boldsymbol{y}}$.

Initialize $\tilde{\mathcal{M}}$, $\mathcal{M}_i$, $\mathcal{C}_i$ according to Eqs. (13) and (14), randomly generate $\boldsymbol{y}_1$ satisfying all constraints, $\mathcal{A}=\{\boldsymbol{y}_1\}$, $t=1$;
while $t<T$ **do**
  $\boldsymbol{\mu}\leftarrow[1/t;\ldots;1/t]\in\Delta^t$;
  while $\boldsymbol{\mu}$ not converge **do**
    Solve $\boldsymbol{\alpha},\boldsymbol{\beta}$ according to Eq. (8);
    Compute $\boldsymbol{w}_1,\ldots,\boldsymbol{w}_t$ according to Eq. (9);
    Solve $\boldsymbol{\mu}$ according to Eq. (10);
  end while
  $\boldsymbol{H}\leftarrow\mathrm{diag}(\boldsymbol{\alpha}-\boldsymbol{\beta})\boldsymbol{K}\mathrm{diag}(\boldsymbol{\alpha}-\boldsymbol{\beta})$;
  $\tilde{\boldsymbol{y}}\leftarrow\mathrm{argmax}_{\hat{\boldsymbol{y}}\in\mathcal{A}}\,\hat{\boldsymbol{y}}^{\mathrm{T}}\boldsymbol{H}\hat{\boldsymbol{y}}$;
  $\tilde{\boldsymbol{y}}\leftarrow\mathrm{sgn}(\boldsymbol{H}\tilde{\boldsymbol{y}})$;
  while $\tilde{\boldsymbol{y}}$ not satisfy all constraints **do**
    For any violated constraint (i,j), compute $\tilde{\boldsymbol{y}}_i$, $\tilde{\boldsymbol{y}}_j$ according to Eq. (15);
    Update $\tilde{\boldsymbol{y}}$ according to Eq. (16);
  end while
  $\mathcal{A}\leftarrow\mathcal{A}\vee\{\tilde{\boldsymbol{y}}\}$;
  $t\leftarrow t+1$;
  if variation of objective value is smaller than τ **then**
    Output $\tilde{\boldsymbol{y}}$, break;
  end if
end while

5 实验

本节通过在 12 个真实数据集上与 6 种半监督聚类算法进行对比, 以验证所提算法的有效性.

5.1 实验设置

数据集. 我们共采用了 12 个真实数据集[1)], 涉及文本、图像、语音等多种数据类型, 其中文本数据有邮件文本、网页文本、软件代码等, 图像包含医学图像、生物图像、遥感图像、手写数字图像等. 表 1 展示了所用数据集的基本信息, 包括数据集名称、样本数和特征数. 样本数规模从 64 到 20000, 特征数规模从 4 到 7129, 涵盖了很大的范围, 具有相当的代表性. 对于每个数据集, 所有特征都被归一化到 $[0,1]$ 之间, 以使得模型的优化过程, 尤其是基于度量学习的模型, 不受部分取值极端的特征的影响.

约束生成. 样本间约束的生成方式则是采用了与文献 [28, 29] 相同的方式, 即随机选取两个样本, 如果它们的真实类别标记相同, 就将其标为 "必连" 约束, 否则标为 "勿连" 约束. 每个数据集上随机生成 100 组约束, 并将其分解为 "必连" 约束集合 $\mathcal{M}$ 和 "勿连" 约束集合 $\mathcal{C}$, 然后根据式 (13) 将 "必连" 约束扩充为 $\tilde{\mathcal{M}}$, 据此对任意样本 $\boldsymbol{x}_i$ 计算与其构成 "必连" 约束关系的全部样本的下标构成的集合 $\mathcal{M}_i$ 和构成 "勿连" 约束关系的全部样本的下标构成的集合 $\mathcal{C}_i$.

评价指标. 本文采用 Rand 指数 [43] (Rand index, RI)、FM 指数 [44] (Fowlkes-Mallows index, FMI)、归一化互信息 (normalized mutual information, NMI) 这 3 个指标来评价所有算法的聚类性能, 其中 Rand 指数是聚类结果与真实类别标记在任意一对样本上预测一致的概率; FM 指数是 "聚类版" 的 F_1

1) https://www.openml.org/search?type=data.

表 1 实验中所用数据集的基本信息

Table 1 Basic statistics of data sets involved in the experiments

ID	Data set	#Instance	#Feature	ID	Data set	#Instance	#Feature
1	dbworld	64	3721	7	pc4	1458	37
2	leukemia	72	7129	8	hivaAgnostic	4229	1617
3	fruitfly	125	4	9	wilt	4839	5
4	semeion	319	256	10	pageBlocks	5473	10
5	pizzaCutter	661	37	11	JapaneseVowels	9961	14
6	pieChart	705	37	12	letter	20000	16

分数 (F_1 score), 即精度 (precision) 和召回率 (recall) 的调和平均; 互信息计算算法给出的聚类结果包含真实聚类结构的信息量, 归一化可以对簇数目较大的聚类结果进行惩罚, 避免得到平凡的结果. 这些性能度量的结果均在 $[0,1]$ 区间, 值越大越好.

对比方法. 本文将 ODMSSC 与第 2.3 小节介绍的 6 种半监督算法进行对比, 包括 3 种使用约束来加速搜索最优聚类结构的算法 COPKMEANS [27], LCVQE [29], PCKMEANS [30]; 2 种先利用约束进行度量学习再进行聚类的算法 MPCKMEANS 和 MKMEANS [33], 其中前者限制了要学习的度量矩阵为对角阵, 为后者的特例; 以及 1 种基于自动编码器的深度约束聚类算法 DCC [41]. 所有算法的簇标记数目设置为真实的类别数目, 其余超参数都通过交叉验证选择得到, 即首先将数据集 $\mathcal{S}$ 划分成 k 个部分, 然后使用 $k-1$ 个部分建立一个聚类模型, 并使用剩下的一部分按所选评价指标验证聚类的效果. 对于任意正整数 k, 依次使用每一部分作为验证集, 重复该过程 k 次, 取 k 次效果的平均值作为该组超参数下的总体效果.

5.2 聚类性能

我们在每个数据集上均进行了 30 次随机实验, 表 2 总结了所有算法在 12 个数据集上 3 种评价指标下的均值和标准差. 每个数据集上的最好结果以粗体显示, •/◦ 代表对应数据集上 ODMSSC 在以 95% 显著性水平的成对 t 检验意义下显著优/劣于对比方法.

表 2 显示, 在 RI, FMI, NMI 这 3 种评价指标下, ODMSSC 分别在 11, 11, 10 个数据集 (总计 12 个数据集) 上取得最优的性能, 这体现了 ODMSSC 的优越性. 此外, 根据最后 3 行对比其他方法的胜/平/负次数来看, OSDMSSC 在大多时候都显著优于其他对比方法, 只在数据集 pc4 上 NMI 指标的结果略低于 LCVQE 和 MKMEANS.

5.3 约束个数的影响

我们还考察了约束个数对聚类性能的影响, 图 1 显示了在 12 个数据集上, 当约束个数从 25 逐渐增加至 100 时, 所有方法的 Rand 指数的变化情况. 从图 1 可以看出, 随着约束个数的增加, ODMSSC 的性能会有所提升, 参考数据集 dbworld, leukemia, fruitfly 的结果, 或在较好的性能处维持相对稳定, 这表明 ODMSSC 在取得良好性能的同时还能保持对超参数的鲁棒性. 与此相对, 其他方法则只能在相对较差的性能处维持稳定, 即使初始只有 25 个约束时性能不好, 随着约束个数的增多性能也鲜有提升, 这表明它们对监督信息的利用效率不高, 甚至还会出现性能下降的情况, 参考数据集 semeion 的结果, 这可能是因为当约束增多时, 这些方法为减少计算开销, 在每轮更新簇标记赋值时多采用了近似或贪心的策略而引入了误差.

表 2 在 12 个真实数据集上的性能比较

Table 2 Performance comparisons on twelve real-world data sets

Data set	Measure	COPKMEANS	LCVQE	PCKMEANS	MKMEANS	MPCKMEANS	DCC	ODMSSC
	RI	.497±.002•	.845±.035•	.496±.000•	.805±.034•	.503±.007•	.850±.011•	**.938±.068**
dbworld	FMI	.623±.057•	.844±.035•	.671±.000•	.806±.034•	.677±.018•	.882±.009•	**.935±.069**
	NMI	.010±.008•	.587±.069•	.003±.000•	.509±.066•	.043±.027•	.623±.012•	**.835±.148**
	RI	.532±.039•	.818±.145•	.579±.008•	.595±.022•	.546±.035•	.847±.042•	**.928±.055**
leukemia	FMI	.625±.054•	.828±.138•	.597±.006•	.611±.020•	.643±.059•	.888±.034•	**.933±.054**
	NMI	.077±.014•	.540±.262•	.122±.009•	.139±.023•	.076±.036•	.815±.062•	**.827±.124**
	RI	.505±.005•	.498±.002•	.502±.002•	.505±.000•	.504±.004•	**.598±.071**	.596±.043
fruitfly	FMI	.545±.046•	.505±.005•	.518±.007•	.522±.000•	.567±.043	**.611±.066**	.607±.041
	NMI	.012±.010•	.005±.004•	.007±.004•	.007±.000•	.007±.009•	**.201±.090**	.197±.084
	RI	.842±.011•	.978±.003•	.977±.000•	.981±.000	.833±.019•	.980±.010	**.984±.004**
semeion	FMI	.840±.010•	.975±.003•	.972±.000•	.980±.000	.830±.018•	.981±.010	**.983±.004**
	NMI	.636±.013•	.910±.010•	.904±.000•	.926±.000	.618±.031•	.923±.045•	**.935±.014**
	RI	.529±.001•	.546±.049•	.532±.000•	.529±.000•	.592±.118•	.793±.011•	**.827±.002**
pizzaCutter	FMI	.682±.001•	.695±.038•	.684±.000•	.686±.000•	.729±.090•	.887±.021•	**.907±.003**
	NMI	.001±.001•	.005±.015•	.001±.001•	.001±.000•	.002±.001•	.016±.009	**.019±.001**
	RI	.536±.001•	.538±.056•	.536±.001•	.531±.000•	.535±.002•	.801±.008•	**.814±.004**
pieChart	FMI	.674±.001•	.673±.044•	.678±.001•	.675±.000•	.673±.002•	.873±.009•	**.899±.002**
	NMI	.001±.001•	.008±.023•	.000±.000•	.001±.000•	.001±.001•	.022±.002•	**.026±.001**
	RI	.507±.000•	.503±.000•	.504±.000•	.505±.000•	.548±.079•	.752±.076•	**.788±.001**
pc4	FMI	.638±.000•	.633±.001•	.638±.000•	.637±.000•	.675±.074•	.861±.083•	**.885±.001**
	NMI	.028±.003	**.032±.001**○	.024±.002•	.032±.000○	.024±.011•	.030±.006	.029±.001
	RI	.512±.001•	.514±.003•	.518±.000•	.513±.001•	.515±.003•	.862±.011•	**.930±.001**
hivaAgnostic	FMI	.692±.001•	.698±.002•	.694±.000•	.693±.000•	.696±.002•	.922±.015•	**.965±.002**
	NMI	.005±.003•	.003±.002•	.002±.000•	.008±.000•	.005±.003•	.013±.009	**.015±.006**
	RI	.500±.000•	.500±.000•	.500±.000•	.500±.000•	.500±.000•	.872±.007•	**.900±.013**
wilt	FMI	.670±.000•	.670±.000•	.670±.000•	.670±.000•	.670±.000•	.908±.012•	**.948±.010**
	NMI	.018±.002•	.018±.000•	.016±.001•	.018±.000•	.018±.001•	.034±.002•	**.040±.002**
	RI	.633±.000•	.634±.000•	.635±.000•	.639±.000•	.638±.000•	.815±.005	**.818±.000**
pageBlocks	FMI	.752±.000•	.757±.000•	.754±.000•	.757±.000•	.758±.000•	.899±.005	**.903±.000**
	NMI	.006±.000•	.007±.000•	.006±.000•	.006±.000•	.006±.000•	.009±.001	**.010±.000**
	RI	.501±.000•	.501±.000•	.501±.000•	.501±.000•	.501±.000•	.724±.004	**.728±.000**
JapaneseVowels	FMI	.604±.000•	.604±.000•	.604±.000•	.604±.000•	.604±.000•	.831±.012•	**.851±.000**
	NMI	.003±.000	.003±.000	.003±.000	.003±.000	.003±.000	.003±.000	**.004±.000**
	RI	.507±.000•	.507±.000•	.507±.000•	.507±.000•	.507±.000•	.810±.023•	**.922±.000**
letter	FMI	.684±.000•	.684±.000•	.684±.000•	.684±.000•	.684±.000•	.910±.035•	**.960±.000**
	NMI	.000±.000•	.000±.000•	.000±.000•	.000±.000•	.000±.000•	.002±.000	**.002±.000**
	RI	12/0/0	12/0/0	12/0/0	11/1/0	12/0/0	8/4/0	
$w/t/l$	FMI	12/0/0	12/0/0	12/0/0	11/1/0	11/1/0	9/3/0	
	NMI	10/2/0	10/1/1	11/1/0	9/2/1	11/1/0	5/7/0	

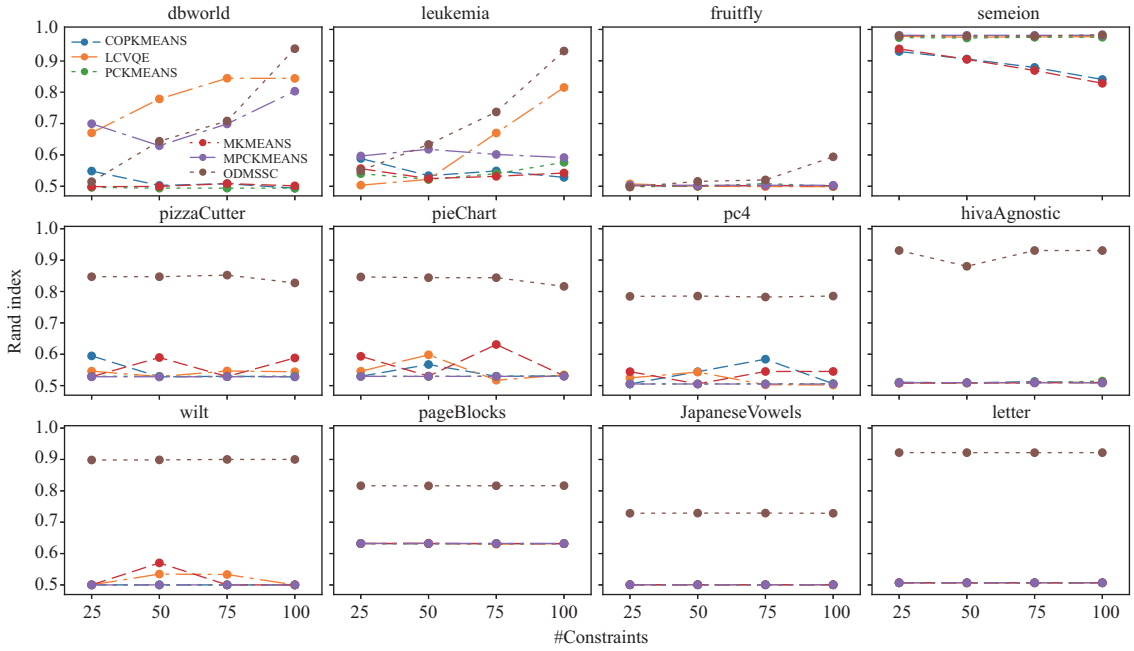

图 1 (网络版彩图) 在不同数据集上约束个数对 Rand 指数的影响

Figure 1 (Color online) The effect of the number of constraints on Rand index

6 结束语

针对带有 “必连” 约束和 “勿连” 约束信息的聚类任务, 本文提出面向半监督聚类的最优间隔分布学习机 (ODMSSC), 旨在通过优化间隔分布和利用监督信息提升聚类性能. ODMSSC 对应的形式化是一个混合整数规划, 本文提出了一种高效的交替优化方法进行求解. 未来, 我们打算扩展 ODMSSC 使其能应用于其他形式的监督信息, 例如聚类簇的尺寸约束等, 同时对其做更深入的理论分析.

参考文献

1 Jain A K, Dubes R C. Algorithms for Clustering Data. Upper Saddle River: Prentice-Hall, 1988

2 Jain A K, Murty M N, Flynn P J. Data clustering: a review. ACM Comput Surv, 1999, 31: 264–323

3 Xu R, WunschII D. Survey of clustering algorithms. IEEE Trans Neural Netw, 2005, 16: 645–678

4 Berkhin P. A survey of clustering data mining techniques. In: Proceedings of Grouping Multidimensional Data: Recent Advances in Clustering, 2006. 25–71

5 Jain A K. Data clustering: 50 years beyond K-means. Pattern Recogn Lett, 2010, 31: 651–666

6 Xu L L, Neufeld J, Larson B, et al. Maximum margin clustering. In: Proceedings of Advances in Neural Information Processing Systems, Vancouver, 2004. 1537–1544

7 Valizadegan H, Jin R. Generalized maximum margin clustering and unsupervised kernel learning. In: Proceedings of Advances in Neural Information Processing Systems, 2006. 1417–1424

8 Li Y F, Tsang I W, Kwok J, et al. Tighter and convex maximum margin clustering. In: Proceedings of the 12th International Conference on Artificial Intelligence and Statistics, Clearwater, 2009. 344–351

9 Zhang K, Tsang I W, Kwok J T. Maximum margin clustering made practical. In: Proceedings of the 24th International Conference on Machine Learning, 2007. 1119–1126

10 Zhao B, Wang F, Zhang C S. Efficient maximum margin clustering via cutting plane algorithm. In: Proceedings of the SIAM International Conference on Data Mining, 2008. 751–762

11 Zhang T, Zhou Z H. Optimal margin distribution clustering. In: Proceedings of the 32nd AAAI Conference on Artificial Intelligence, 2018. 4474–4481
12 Gao W, Zhou Z H. On the doubt about margin explanation of boosting. Artif Intell, 2013, 203: 1–18
13 Basu S, Davidson I, Wagstaff K. Constrained Clustering: Advances in Algorithms, Theory, and Applications. Boca Raton: Chapman & Hall/CRC, 2008
14 Cristianini N, Shawe-Taylor J. An Introduction to Support Vector Machines and Other Kernel-based Learning Methods. Cambridge: Cambridge University Press, 2000
15 Cortes C, Vapnik V N. Support-vector networks. Mach Learn, 1995, 20: 273–297
16 Vapnik V N. The Nature of Statistical Learning Theory. New York: Springer, 1995
17 Zhou Z H. Large margin distribution learning. In: Proceedings of the 6th IAPR International Workshop on Artificial Neural Networks in Pattern Recognition, Montreal, 2014. 1–11
18 Zhang T, Zhou Z H. Optimal margin distribution machine. IEEE Trans Knowl Data Eng, 2020, 32: 1143–1156
19 Smola A J, Schölkopf B. A tutorial on support vector regression. Stat Comput, 2004, 14: 199–222
20 Sion M. On general minimax theorems. Pac J Math, 1958, 8: 171–176
21 Gönen M, Alpaydın E. Multiple kernel learning algorithms. J Mach Learn Res, 2011, 12: 2211–2268
22 Zhang K, Tsang I W, Kwok J T. Maximum margin clustering made practical. IEEE Trans Neural Netw, 2009, 20: 583–596
23 Yuille A L, Rangarajan A. The concave-convex procedure. Neural Comput, 2003, 15: 915–936
24 Zhou G T, Lan T, Vahdat A, et al. Latent maximum margin clustering. In: Proceedings of Advances in Neural Information Processing Systems, Lake Tahoe, 2013. 28–36
25 Niu G, Dai B, Shang L, et al. Maximum volume clustering: a new discriminative clustering approach. J Mach Learn Res, 2013, 14: 2641–2687
26 Vijaya Saradhi V, Charly Abraham P. Incremental maximum margin clustering. Pattern Anal Applic, 2016, 19: 1057–1067
27 Wagstaff K, Cardie C, Rogers S, et al. Constrained k-means clustering with background knowledge. In: Proceedings of the 18th International Conference on Machine Learning, Williamstown, 2001. 577–584
28 Davidson I, Ravi S. Clustering with constraints: feasibility issues and the k-means algorithm. In: Proceedings of SIAM International Conference on Data Mining, Newport Beach, 2005. 138–149
29 Pelleg D, Baras D. K-means with large and noisy constraint sets. In: Proceedings of the 18th European Conference on Machine Learning, Warsaw, 2007. 674–682
30 Basu S, Banerjee A, Mooney R J. Active semi-supervision for pairwise constrained clustering. In: Proceedings of SIAM International Conference on Data Mining, Lake Buena Vista, 2004. 333–344
31 Lu Z D, Leen T K. Pairwise constraints as priors in probabilistic clustering. In: Proceedings of Constrained Clustering: Advances in Algorithms, Theory, and Applications, 2008. 59–90
32 Shental N, Bar-Hillel A, Hertz T, et al. Gaussian mixture models with equivalence constraints. In: Proceedings of Constrained Clustering: Advances in Algorithms, Theory, and Applications, 2008. 33–58
33 Xing E P, Ng A Y, Jordan M I, et al. Distance metric learning with application to clustering with side-information. In: Proceedings of Advances in Neural Information Processing Systems, Vancouver, 2002. 505–512
34 Hertz T, Bar-Hillel A, Weinshall D. Boosting margin based distance functions for clustering. In: Proceedings of the 21st International Conference on Machine Learning, 2004. 50–57
35 Freund Y, Schapire R E. A decision-theoretic generalization of on-line learning and an application to boosting. In: Proceedings of the 2nd European Conference on Computational Learning Theory, Barcelona, 1995. 23–37
36 Liu Y, Jin R, Jain A K. Boostcluster: boosting clustering by pairwise constraints. In: Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Jose, 2007. 450–459
37 Yi J F, Zhang L J, Yang T B, et al. An efficient semi-supervised clustering algorithm with sequential constraints. In: Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Sydney, 2015. 1405–1414
38 Cohn D, Caruana R, McCallum A. Semi-supervised clustering with user feedback. In: Proceedings of Constrained Clustering: Advances in Algorithms, Theory, and Applications, 2003. 17–32
39 Settles B. Active learning. In: Proceedings of Synthesis Lectures on Artificial Intelligence and Machine Learning, 2012

40 Goodfellow I, Bengio Y, Courville A. Deep Learning. Cambridge: MIT Press, 2016
41 Zhang H, Basu S, Davidson I. A framework for deep constrained clustering — algorithms and advances. In: Proceedings of Joint European Conference on Machine Learning and Knowledge Discovery in Databases, 2020. 57–72
42 Kim S J, Boyd S. A minimax theorem with applications to machine learning, signal processing, and finance. In: Proceedings of the 46th IEEE Conference on Decision and Control, New Orleans, 2007. 1344–1367
43 Rand W M. Objective criteria for the evaluation of clustering methods. J Am Stat Assoc, 1971, 66: 846–850
44 Fowlkes E B, Mallows C L. A method for comparing two hierarchical clusterings. J Am Stat Assoc, 1983, 78: 553–569

Optimal margin distribution machine for semi-supervised clustering

Teng ZHANG[1,2,3,4*], Ming LI[5] & Hai JIN[1,2,3,4]

1. *National Engineering Research Center for Big Data Technology and System, Huazhong University of Science and Technology, Wuhan* 430074*, China*;
2. *Service Computing Technology and System Lab, Huazhong University of Science and Technology, Wuhan* 430074*, China*;
3. *Cluster and Grid Computing Lab, Huazhong University of Science and Technology, Wuhan* 430074*, China*;
4. *School of Computer Science and Technology, Huazhong University of Science and Technology, Wuhan* 430074*, China*;
5. *National Key Laboratory for Novel Software Technology, Nanjing University, Nanjing* 210023*, China*
* Corresponding author. E-mail: tengzhang@hust.edu.cn

Abstract Margin-based clustering is one of the most classical clustering algorithms, which assumes that the best clustering structure can be determined by introducing margin used in supervised learning. That is for a satisfactory clustering result, when used as labels for supervised learning, some margin-related statistics produced by the obtained classifier can simultaneously be optimal. Currently, the most optimal statistic is the margin distribution, which bases on the latest margin theory and has achieved better results than optimizing the minimum margin. However, in some real clustering tasks, there is extra supervised information available such as the "must-link" and "cannot-link" constraints between a pair of instances, and the effectiveness of optimizing margin distribution in these circumstances has not been well exploited. In this paper, we propose an optimal margin distribution machine for semi-supervised clustering (ODMSSC), whose formulation is mixed-integer programming. We adopt the minimax convex relaxation to convert it into a saddle point problem, and propose an efficient alternating optimization method to solve the problem. Extensive experiments on real data sets also verify the superiority of the proposed method.

Keywords semi-supervised clustering, constrained clustering, optimal margin distribution machine, margin distribution, margin