谱聚类算法的泛化性能分析
On the Generalization Ability of Spectral Clustering Algorithms

报告人：梁伟轩

国防科技大学计算机学院

2021 年 6 月 4 日

提纲

提纲

设样本集为 $\{\mathbf{x}_i\}_{i=1}^n$，独立同分布．假设通过函数 $W(\cdot,\cdot)$ 构建的边集为 $\mathbf{W}\in\mathbb{R}^{n\times n}$，其中元素 $\mathbf{W}_{ij}=\frac{1}{n}W(\mathbf{x}_i,\mathbf{x}_j)$．

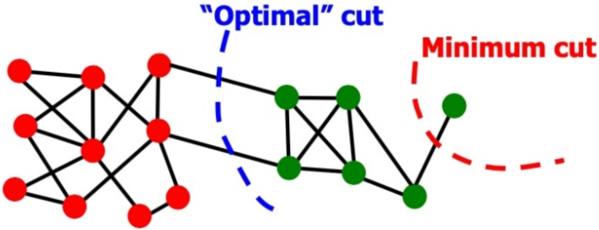

图: 谱聚类示意图

下面分别介绍两类谱聚类的方式:

$$RatioCut(\mathbb{V}_1,..,\mathbb{V}_K)=\frac{1}{2}\sum_{k=1}^{K}\frac{Cut(\mathbb{V}_k,\overline{\mathbb{V}}_k)}{|\mathbb{V}_k|}\qquad NCut(\mathbb{V}_1,...,\mathbb{V}_K)=\frac{1}{2}\sum_{k=1}^{K}\frac{Cut(\mathbb{V}_k,\overline{\mathbb{V}}_k)}{vol(\mathbb{V}_k)}$$

令 $\mathbf{U}=(\mathbf{u}_1,\dots,\mathbf{u}_K)\in\mathbb{R}^{n\times K}$ 为聚类指示函数，对于该矩阵每一行，有且仅有一个元素不为 0．定义经验风险函数为

$$\hat{F}(\mathbf{U}):=\frac{1}{2n(n-1)}\sum_{k=1}^{K}\sum_{i\neq j}^{n}\mathbf{W}_{ij}(\mathbf{u}_{i,k}-\mathbf{u}_{j,k})^2$$

1. 若为 RatioCut，那么聚类指示矩阵满足如下条件：
当 $\mathbf{x}_i\in\mathbb{V}_k$ 时，有 $\mathbf{u}_{i,k}=\frac{1}{\sqrt{|\mathbb{V}_k|}}$；当 $\mathbf{x}_i\notin\mathbb{V}_k$ 时，则 $\mathbf{u}_{i,k}=0$．

2. 若为 NCut，那么聚类指示矩阵满足如下条件：
当 $\mathbf{x}_i\in\mathbb{V}_k$ 时，有 $\mathbf{u}_{i,k}=\frac{1}{\sqrt{vol(\mathbb{V}_k)}}$；当 $\mathbf{x}_i\notin\mathbb{V}_k$ 时，则 $\mathbf{u}_{i,k}=0$．

上面两个问题由于离散约束，与 k 均值聚类一样，是 NP 难的问题，所以对聚类指示矩阵进行松弛．

若为 RatioCut，则优化式化为：

$$\min_{\mathbf{U}} \hat{F}(\mathbf{U}),\ \ s.t.\ \mathbf{U}^\top\mathbf{U} = \mathbf{I}_K$$

若为 NCut，则优化式化为：

$$\min_{\mathbf{U}} \hat{F}(\mathbf{U}),\ \ s.t.\ \mathbf{U}^\top\mathbf{D}\mathbf{U} = \mathbf{I}_K$$

其中 $\mathbf{D}$ 为度矩阵，是由 $\mathbf{W}$ 各行行和组成的对角矩阵.

可将 $\hat{F}(\mathbf{U})$ 写成矩阵形式：$\hat{F}(\mathbf{U}) = \frac{1}{n(n-1)}\mathrm{tr}(\mathbf{U}^\top\mathbf{L}\mathbf{U})$，其中 $\mathbf{L} = \mathbf{D} - \mathbf{W}$.
上式的解为特征分解问题.

提纲

期望风险 (Expected Risk)

定义经验风险函数的极限形式:

$$F(U) = \frac{1}{2}\sum_{k=1}^{K}\iint W(x,y)(u_k(x)-u_k(y))^2\,d\rho(x)\rho(y)$$

相应的 RatioCut 问题有如下形式:

$$\min_{U} F(U),\ \ s.t.\ \ u_i(x) = 1/\sqrt{\int_{\mathbb{V}_i} d\rho(x)},\ \ if\ x\in\mathbb{V}_i,\ \ otherwise\ 0.$$

对应地进行松弛操作，有如下形式:

$$\min_{U} F(U),\ \ s.t.\ \ <u_i, u_j>_\rho = 1\ \ if\ i=j,\ \ otherwise\ 0.$$

期望风险 (Expected Risk)

设积分算子 $(L_K f)(x) = \int L(x, y) f(y)\, d\rho(y)$. 据此，可以将 $F(U)$ 化为

$$F(U) = \sum_{k=1}^{K} \int u_k(x) L_K u_k(x)\, d\rho(x)$$

其中 $L(x, y) = m(x) - W(x, y)$, $m(x) = \int W(x, y)\, d\rho(y)$. 该式的最优解即为算子 L_K 最小的 K 个特征值对应的特征函数，设为 $\widetilde{U}^* = (\widetilde{u}_1^*, ..., \widetilde{u}_K^*)$.

下面重点研究，如何计算经验风险最小的解的期望风险. 设 $\widetilde{\mathbf{U}}^* = (\widetilde{\mathbf{u}}_1^*, ..., \widetilde{\mathbf{u}}_K^*) \in \mathbb{R}^{n\times K}$ 为经验风险最小的一组解. 定义算子 $T_n : \mathcal{H} \to \mathcal{H}$ 如下:

$$T_n = \frac{1}{n} \sum_{i=1}^{n} < \cdot, L_{\mathbf{x}_i} > L_{\mathbf{x}_i}$$

其中 $L_{\mathbf{x}_i} = L(\cdot, \mathbf{x}_i)$. 那么 T_n 与 $\mathbf{L}$ 有相同的特征值.

对于特征值 λ_k，T_n 对应的特征函数为 $\check{u}_k(x)$，$\mathbf{L}$ 对应的特征向量为 $\widetilde{\mathbf{u}}_k^*$. 那么它们有如下对应关系：

$$\widetilde{\mathbf{u}}_k^* = \frac{1}{\sqrt{\lambda_k}}(\check{u}_k(\mathbf{x}_1), \ldots, \check{u}_k(\mathbf{x}_n)); \quad \check{u}_k(x) = \frac{1}{\sqrt{\lambda_k}}\left(\frac{1}{n}\sum_{i=1}^{n} \widetilde{\mathbf{u}}_{i,k}^* L(x, \mathbf{x}_i)\right)$$

所以需要界定如下式子的上界，亦被称为期望超出风险 (Expected Excess Risk)：

$$F(\check{U}) - F(\widetilde{U}^*)$$

提纲

谱聚类简介

风险分析问题的引入

主要定理 1

定理 1 的证明

定理 1

假设对于任意的 $\check{u} \in \mathcal{H}$，有 $\|\check{u}\|_\infty \le \sqrt{B}$，那么对于任意的 $\delta > 0$，至少有 $1-2\delta$ 的概率，下式成立：

$$F(\check{U}) - F(\widetilde{U}^*) \le 8CBK\left(\sqrt{\frac{1}{n}} + 2\sqrt{\frac{2\log\frac{1}{\delta}}{n}}\right) + K\frac{2\kappa\sqrt{2\log\frac{2}{\delta}}}{\sqrt{n}} \tag{1}$$

其中 C 和 B 为正常数，K 为聚类数，$\kappa = \sup_{x\in\mathcal{X}} L(x,x)$.

证明思路，将 $F(\check{U}) - F(\widetilde{U}^*)$ 拆分为：

$$F(\check{U}) - F(\widetilde{U}^*) = \underbrace{F(\check{U}) - \hat{F}(\widetilde{\mathbf{U}}^*)}_{\mathcal{B}} + \underbrace{\hat{F}(\widetilde{\mathbf{U}}^*) - F(\widetilde{U}^*)}_{\mathcal{C}}$$

然后分别给出两部分的上界.

提纲

令 $l_u(x,y) = W(x,y)(u(x)-u(y))^2$，及 $\check{u}'_k(x) = \frac{1}{\sqrt{\lambda}}\check{u}_k(x)$. 首先给出第一部分的上界：

$$\begin{aligned}
\mathcal{B} &= F(\check{U}) - \hat{F}(\widetilde{\mathbf{U}}^*) \\
&= \frac{1}{2}\sum_{k=1}^{K}\left[\iint W(x,y)(\check{u}_k(x)-\check{u}_k(y))^2 d\rho(x)d\rho(y) - \frac{1}{n(n-1)}\sum_{i\neq j}^{n} W(\mathbf{x}_i,\mathbf{x}_j)(\widetilde{\mathbf{u}}^*_{i,k}-\widetilde{\mathbf{u}}^*_{j,k})^2\right] \\
&= \frac{1}{2}\sum_{k=1}^{K}\left[\iint W(x,y)(\check{u}_k(x)-\check{u}_k(y))^2 d\rho(x)d\rho(y) - \frac{1}{n(n-1)}\sum_{i\neq j}^{n} W(\mathbf{x}_i,\mathbf{x}_j)\left(\frac{1}{\sqrt{\lambda_k}}\check{u}_k(\mathbf{x}_i)-\frac{1}{\sqrt{\lambda_k}}\check{u}_k(\mathbf{x}_j)\right)^2\right] \\
&= \frac{1}{2}\sum_{k=1}^{K}\left[\iint W(x,y)(\check{u}_k(x)-\check{u}_k(y))^2 d\rho(x)d\rho(y) - \frac{1}{n(n-1)}\sum_{i\neq j}^{n} W(\mathbf{x}_i,\mathbf{x}_j)(\check{u}'_k(x_i)-\check{u}'_k(x_j))^2\right] \\
&= \frac{1}{2}\sum_{k=1}^{K}\left\{E[l_{\check{u}_k}] - \hat{E}[l_{\check{u}'_k}]\right\}
\end{aligned}$$

进一步，由于

$$E[l_{\breve{u}_k}] - \hat{E}[l_{\breve{u}'_k}] \leq E[l_{\breve{u}_k}] - \hat{E}[l_{\breve{u}_k}] \\ \leq \sup_{u\in\mathcal{H}}(E[l_u] - \hat{E}[l_u])$$

所以，$\mathcal{B} \leq \frac{1}{2}K\sup_{u\in\mathcal{H}}(E[l_u] - \hat{E}[l_u])$. 可以利用 McDiarmid 不等式，以 $1-\delta$ 的概率，有

$$\sup_{u\in\mathcal{H}}(E[l_u] - \hat{E}[l_u]) \leq E\sup_{u\in\mathcal{H}}(E[l_u] - \hat{E}[l_u]) + 32CB\sqrt{\frac{2\log\frac{1}{\delta}}{n}}$$

此外，上式中右边第一项表示为：

$$\sup_{u\in\mathcal{H}}(E[l_u] - \hat{E}[l_u]) \leq E\sup_{u\in\mathcal{H}}\left(E[l_u] - \frac{1}{\lfloor\frac{n}{2}\rfloor}\sum_{i=1}^{\lfloor\frac{n}{2}\rfloor} l_u(\mathbf{x}_i, \mathbf{x}_{i+\lfloor\frac{n}{2}\rfloor})\right)$$

进一步，利用关于 Rademacher 复杂度的讨论，可以得到如下上界：

$$E\sup_{u\in\mathcal{H}}(E[l_u] - \hat{E}[l_u]) \leq 16CB\sqrt{\frac{1}{n}}$$

综上，可以得到 $\mathcal{B} \leq 8CBK(\sqrt{\frac{1}{n}} + 2\sqrt{\frac{2\log\frac{1}{\delta}}{n}})$ 以 $1-\delta$ 的概率成立.

下面来计算 $\mathcal{C}$ 的上界. $F(\widetilde{U}^*)$ 可以化为如下形式:

$$\frac{1}{2}\sum_{k=1}^{K}\iint W(x,y)(\widetilde{u}_k^*(x)-\widetilde{u}_k^*(y))^2\,d\rho(x)\,d\rho(y)$$

$$\sum_{k=1}^{K}\int \widetilde{u}_k^*(x)L_K\widetilde{u}_k^*(x)\,d\rho(x)=\sum_{k=1}^{K}<\widetilde{u}_k^*,L_k\widetilde{u}_k^*>=\sum_{k=1}^{K}\lambda_k(L_K)<\widetilde{u}_k^*,\widetilde{u}_k^*>=\sum_{k=1}^{K}\lambda_k(L_K)$$

进一步，可知:

$$\begin{aligned}\mathcal{C}=\hat{F}(\widetilde{\mathbf{U}}^*)-F(\widetilde{U}^*)&=\sum_{k=1}^{K}(\lambda_k(\mathbf{L})-\lambda_k(L_K))\\ &\le K\sup_k|\lambda_k(T_n)-\lambda_k(T_{\mathcal{H}})|\le K\|T_n-T_{\mathcal{H}}\|\\ &\le K\|T_n-T_{\mathcal{H}}\|_{HS}\le K\frac{2\sqrt{2}\kappa\sqrt{\log\frac{2}{\delta}}}{\sqrt{n}}\end{aligned}$$

其中，$T_{\mathcal{H}}=\int<\cdot,L_x>L_x d\rho(x)$.

请各位老师同学批评指正

谢　　谢！