面向大数据的扁平聚类算法研究

任远航
201721220117
ryuanhang@gmail.com

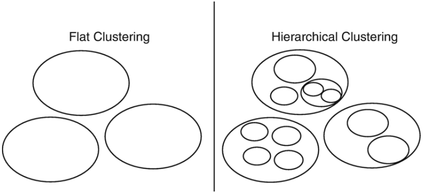
信息与软件工程学院
电子科技大学

2020年5月8日

目录

聚类

- 什么是聚类?
- 扁平聚类vs层级聚类



研究的问题

大数据下满足下列条件的算法是急需的:

- 有理论保证
- 高效

本论文将围绕*k*-means和谱聚类展开研究

目录

目录

k-MEANS问题

- 得益于Lloyd算法[Lloyd, 1982]，这一问题非常有名
- k-means问题是NP难的

定义 (k-MEANS问题)

给定数据集$\mathcal{X} \subseteq \mathbb{R}^d$以及$k$个点的集合$C \subseteq \mathbb{R}^d$，定义如下目标函数

$$\phi_C(\mathcal{X}) = \sum_{x \in \mathcal{X}} d^2(x, C) \tag{1}$$

其中$d(x, C) = \min_{c \in C} \|x - c\|$是点到集合的距离，*k-means*问题的目标是找到最优的C从而使得上式最小

解的质量

定义 (解的质量1)

令$\alpha \geq 1$，如果下式成立，称C是*k-means*问题的一个α近似解

$$\phi_C(\mathcal{X}) \leq \alpha \phi_{OPT}(\mathcal{X}) \tag{2}$$

其中$\phi_{OPT}(\mathcal{X})$是最优距离平方和

定义 (解的质量2)

令$\alpha \geq 1$且$\beta > 0$，如果下式成立，称C是*k-means*问题的一个β-*bad* α近似解

$$\phi_C(\mathcal{X}) > (\alpha + \beta) \phi_{OPT}(\mathcal{X}) \tag{3}$$

否则，C被称为β-*good* α近似解

目录

k-means++ [Arthur and Vassilvitskii, 2007]利用d^2 *weighting*得到了一个$O(\log k)$的近似解

算法 1: k-means++ seeding

输入: 数据集$\mathcal{X}$，类数目k

输出: k个点C

$c_1 \leftarrow$ 从$\mathcal{X}$中均匀采样一个点

$C \leftarrow \{c_1\}$

for $i = 2, 3, \ldots k$ **do**

  for $x \in \mathcal{X}$ **do**

    $p(x) \leftarrow \mathrm{d}(x, C)^2 / \sum_{x' \in \mathcal{X}} \mathrm{d}(x', C)^2$

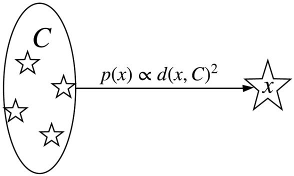
  end

  $x \leftarrow$依分布p从$\mathcal{X}$中采样一个点

  $C \leftarrow C \cup \{x\}$

end

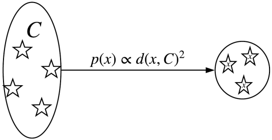
返回 C



k-means|| [Bahmani et al., 2012]加速了k-means++

算法 2: k-means|| seeding

输入: 数据集$\mathcal{X}$，每一轮期望采样数l，聚类数k，轮数t
输出: k个点C
$S \leftarrow$ 从$\mathcal{X}$中均匀采样一个点
for $i = 1,2,\ldots,t$ **do**
  $C' \leftarrow \emptyset$
  for $x \in \mathcal{X}$ **do**
    以概率$\min(1, \frac{ld^2(x,S)}{\phi(\mathcal{X},S)})$将$x$加入$C'$中
  end
  $S \leftarrow S \cup C'$
end
对点$s \in S$，令w_s是$\mathcal{X}$中离s最近的点的数目
$C \leftarrow$ 令w_s是s的权重，在带权的S上运行一个α近似算法
返回 C



第一个贡献

将经典的k-means++算法扩展到了带权的k-means问题上并证明了扩展算法的聚类质量。

带权k-MEANS问题和新算法

定义 (带权k-MEANS问题)

给定数据集$\mathcal{X} \subseteq \mathbb{R}^d$和对应的权重$w$，找到一个大小为$k$的集合$C \subseteq \mathbb{R}^d$使得下面的目标函数最小。

$$\psi_C(\mathcal{X}) = \sum_{x_i \in \mathcal{X}} w_i d^2(x_i, C) \tag{4}$$

算法 3: 带权k-means++ seeding

输入: 数据集$\mathcal{X}$，权重w，类数目k
输出: k个点C
$c_1 \leftarrow$ 依概率$\frac{w_x}{\sum_{i \in \mathcal{X}} w_i}$从$\mathcal{X}$中采样一个点
$C \leftarrow \{c_1\}$
for $i = 2, 3, \dots k$ **do**
  for $x \in \mathcal{X}$ **do**
    $p(x) \leftarrow w_x \mathrm{d}(x, C)^2 / \sum_{x' \in \mathcal{X}} w'_x \mathrm{d}(x', C)^2$
  end
  $x \leftarrow$ 依分布p从$\mathcal{X}$中采样一个点
  $C \leftarrow C \cup \{x\}$
end
返回 C

带权k-MEANS++的理论保证

定理 (带权k-MEANS++的理论保证)

给定数据集$\mathcal{X}$和权重w，令C是带权k-$means$++返回的结果，则

$$\mathbb{E}[\psi_C(\mathcal{X})] \leq 8(\ln k + 2)\psi_{OPT}(\mathcal{X}) \tag{5}$$

证明的框架:

- 考虑均匀采样的解的质量
- 考虑d^2 weighting的质量
- 用归纳法揭示中间解的关系

证明的框架

引理 (均匀采样的解的质量)

考虑任意最优类A，令Z是在A中以概率$\frac{w_z}{\sum_{i\in A} w_i}$被挑到的点，
则，$\mathbb{E}[\psi_Z(A)] = 2\psi_{OPT}(A)$

引理 (d^2 WEIGHTING的质量)

令C是任意中间结果，$1 \le |C| \le k-1$，令A是任意最优类，并且令$Z \in A$是基于C用d^2 *weighting*挑到的点，
则，$\mathbb{E}[\psi_{C'}(A)|C, Z \in A] \le 8\psi_{OPT}(A)$，其中$C' = C \cup \{Z\}$

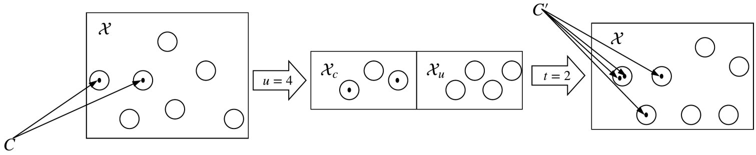
证明的框架

引理 (中间解的关系)

令C是任意中间结果，$1 \leq |C| \leq k-1$，任选$u > 0$个 "未覆盖"的最优类，令$\mathcal{X}_u$是那些类中的点，并且令$\mathcal{X}_c = \mathcal{X} - \mathcal{X}_u$，假定我们基于$C$用$d^2$ *weighting*的方法添加$0 \leq t \leq u$个点，令C'是添加后的解，则，

$$\mathbb{E}[\psi_{C'}(\mathcal{X})|C] \leq [\psi_C(\mathcal{X}_c) + 8\psi_{OPT}(\mathcal{X}_u)](1 + H_t) + \frac{u-t}{u}\psi_C(\mathcal{X}_u) \qquad (6)$$

其中$H_t = 1 + \frac{1}{2} + ... + \frac{1}{t}$是调和数



目录

基于均匀采样的聚类

算法 4: 基于均匀采样的聚类

输入: 数据集$\mathcal{X}$，类数目k，采样数s，聚类算法$\mathcal{A}_c$
输出: k个点C
$S \leftarrow$ 均匀不放回采样s个点
$C \leftarrow$ 在S上用$\mathcal{A}_c$解决k-means问题
返回 k个点C

定理 (算法4的解的质量)

令$0 < \delta < 1/2$，$\alpha \geq 1$，$\beta > 0$是近似系数，令C是均匀采样返回的解且$\mathcal{A}_c$是一个α近似的算法，假定我们均匀不放回采样s个点且

$$s \geq \ln(\frac{1}{\delta})(1 + \frac{1}{n})/(\frac{\beta^2 m^2}{2\Delta^2\alpha^2} + \frac{\ln(1/\delta)}{n}) \tag{7}$$

则，下式以至少$1 - 2\delta$的概率成立，

$$\phi_C(\mathcal{X}) \leq 4(\alpha + \beta)\phi_{OPT}(\mathcal{X}) \tag{8}$$

其中$\Delta = \max_{i,j} \|v_i - v_j\|^2$是数据直径的平方，$m = \phi_{OPT}(\mathcal{X})/n$是最优目标函数值的平均值

第二个贡献

- 对均匀采样给出了一个更紧的理论界，并且证明在温和的数据假设下算法的运行时间在多项式对数级别
- 提出了新的加速算法Double-K-MC^2（估计权重）
- 给出了均匀采样、K-MC^2、Double-K-MC^2，和它们对应的kernel版本的MATLAB实现，用实验验证了这些算法的效率和有效性

一个更紧的理论界

定理 (均匀采样的更紧理论界)

令$0 < \delta < 1/2$，$\alpha \geq 1$，$\beta > 0$是近似系数，令C是均匀采样返回的解且$\mathcal{A}_c$是一个α近似的算法，假定我们均匀不放回采样s个点且

$$s \geq \ln(\frac{1}{\delta})(1+\frac{1}{n})/(\frac{\beta^2 m^2}{2\Delta^2\alpha^2}+\frac{\ln(1/\delta)}{n})$$

则，下式以至少$1-2\delta$的概率成立，

$$\phi_C(\mathcal{X}) \leq (\alpha+\beta)\phi_{OPT}(\mathcal{X})$$

其中$\Delta = \max_{i,j} \|v_i - v_j\|^2$是数据直径的平方，$m = \phi_{OPT}(\mathcal{X})/n$是最优目标函数值的平均值

证明的框架：

1. 说明C对于S会是一个好的解
2. 如果C对于$\mathcal{X}$是一个坏的解，则大概率C对于S会是一个坏的解
3. 根据1和2，说明C对于$\mathcal{X}$是一个好的解

一个多项式对数时间的算法

假设数据是独立的从同一个分布F中采样出来的

- 分布F的尾部是指数衰减的，即$\exists c, t$使得$P[d(x, \mu(F)) > a] \le ce^{-at}$，其中$x \sim F$
- F在一个球面上的最大和最小概率密度的比值被一个大于等于1的常数限制（球面上的密度大于0）

定理 (均匀采样的效率)

令$0 < \delta < 1/2$，$\alpha \ge 1$，$\beta > 0$是近似系数，假定假设*(A1)*和*(A2)*成立，令C是均匀采样返回的解，则，下式以至少$1-2\delta$的概率成立

$$\phi_C(\mathcal{X}) \le (\alpha + \beta)\phi_{OPT}(\mathcal{X})$$

如果采样$O(\ln(\frac{1}{\delta})\frac{\alpha^2}{\beta^2}k^2\log^4 n)$个点的话

目录

基准算法

- 由于均匀采样既高效又有理论保证，我们设计实验来检验这一点
- 基准算法是K-MC^2 [Bachem et al., 2016]和Double-K-MC^2
- 由于k-means$||$中的权重计算很花时间，我们提出一个新的称为Double-K-MC^{2}采样的算法来加速（估算权重）

算法 5: Double-K-MC^{2}采样

输入: 数据集$\mathcal{X}$，采样数s，游走步数u
输出: k个点C
$S_1 \leftarrow$ 在$\mathcal{X}$中用K-MC^{2}采样s个点
$\mathcal{X}' \leftarrow$ 在$\mathcal{X}$中移除S_1
$S_2 \leftarrow$ 在$\mathcal{X}'$中用K-MC^{2}采样s个点
对S_1中任意点s_i，令w_i是S_2中离s_i最近的点的数目
$C \leftarrow$ 令$w_i + 1$是s_i的权重，在带权S_1上运行一个α近似算法
返回 k个点C

传统聚类

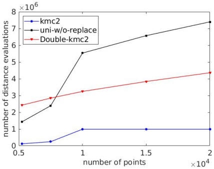
表 1: 数据量n，类数目k，维度d

数据集	n	k	d
a2	5250	35	2
a3	7500	50	2
b2-random-10	10000	100	2
b2-random-15	15000	100	2
b2-random-20	20000	100	2
KDD	145751	200	74
RNA	488565	200	8
Poker Hand	1000000	200	10

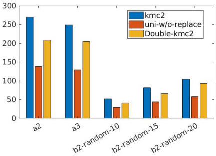
- 游走步数：$u = 200$
- 采样量：Double-K-MC^{2}和均匀采样分别是$1.5 \log^2 n$和$0.7 \log^4 n$
- α近似算法：(带权)k-means++接Lloyd
- 评价指标：距离计算次数和k-means目标函数

结果

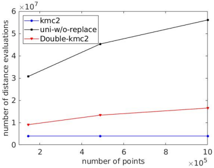
1. 均匀采样差不多比K-MC^{2}慢10倍，不过随着数据的增多时间增长较慢，目标函数值差不多是K-MC^{2}的60%
2. Double-K-MC^{2}比K-MC^{2}聚类质量好时间花费比均匀采样少
3. 如果你想要一个好的聚类质量且时间花费要合理的话，选择Double-K-MC^2，如果追求质量，推荐均匀采样



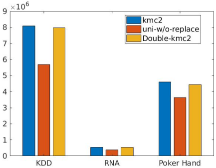
(A) 合成数据集上的距离计算次数



(B) 合成数据集上的k-means目标函数值



(C) 真实数据集上的距离计算次数



(D) 真实数据集上的k-means目标函数值

图 1: k-means目标函数值和时间花费随数据量变化图

图像分割

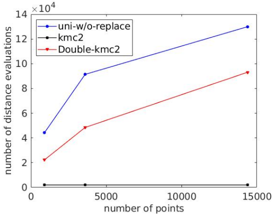
表 2: 数据量n，类数目k

数据集	n	k
baby	900(30 * 30)	5
kitten	3600(60 * 60)	5
bear	14400(120 * 120)	5

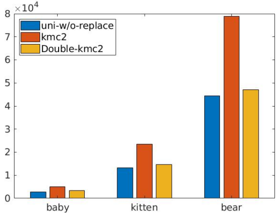
- 均匀采样、Double-K-MC^2，和K-MC^{2}的kernel版
- 用Stella and Shi [2003]的方法计算相似度矩阵A，寻找离A最近的正定矩阵K做为kernel
- 游走步数：$u = 200$
- 采样量：Double-K-MC^{2}和均匀采样分别是$0.25 \log^2 n$和$0.4 \log^4 n$
- α近似算法：(带权)kernel k-means++接kernel Lloyd
- 评价指标：距离计算次数和kernel k-means目标函数值

结果

1. 均匀采样的kernel版有最优的聚类质量而时间开销增长不快
2. Double-K-MC^{2}的kernel版有着和均匀采样差不多的聚类质量而时间开销少很多
3. 因此，如果追求聚类质量，推荐使用Double-K-MC^2，如果要更快的速度，K-MC^{2}是一个更好的选择



(A) 图片数据上的距离计算次数



(B) 图片数据上的kernel k-means目标函数值

图 2: kernel k-means目标函数值和时间开销随数据量变化图

目录

目录

谱聚类
- 背景
- 有理论保证且高效的算法
- 实验

谱聚类问题

- 此问题的来源之一是图论
- 该问题也是NP难的

定义 (Normalized Cut)

记v_i是节点i，C_j是类j，找到一个划分矩阵$F \in \mathbb{R}^{n\times k}$使得下面的目标函数能最小

$$\min_F \operatorname{Tr}(\frac{F^T LF}{F^T DF}) \tag{9}$$

其中$F_{ij} = \begin{cases} 1 & v_i \in C_j \\ 0 & otherwise \end{cases}$，$L = D - A$，$A$是图的相似度矩阵，$D$是$A$的度矩阵

目录

转变为传统聚类

定义 (带权KERNEL k-MEANS问题)
给定数据集$\mathcal{X} \subseteq \mathbb{R}^d$和对应权重$w$以及一个映射函数$\varphi(.)$，找到一个大小为$k$的集合$C$使得下面的目标函数最小

$$\Psi_C(\mathcal{X}) = \sum_{i=1}^{k} \sum_{x_j \in \pi_i} w_j \left\| \varphi(x_j) - C_i \right\|^2 \tag{10}$$

其中π_i表示第i个类，$\cup_{i=1}^{k} \pi_i = \mathcal{X}$，$C_i = \frac{\sum_{x_j \in \pi_i} w_j \varphi(x_j)}{\sum_{j \in \pi_i} w_j}$

引理 (两种聚类间的关系[DHILLON ET AL., 2004])
令W和K是带权*kernel k-means*问题的权重和核矩阵，其中$W = D$，$K = D^{-1}AD^{-1}$，其中D和A是谱聚类的度矩阵和相似度矩阵，则，

$$\Psi_C(\mathcal{X}) = \mathrm{Tr}(D^{-1/2}AD^{-1/2}) - k + Ncut \tag{11}$$

基于均匀采样的算法

算法 6: 基于均匀采样和带权kernel k-means的谱聚类算法[Mohan and Monteleoni, 2017]

输入: 数据集$\mathcal{X}$，类数目k，采样数s，相似度矩阵A，度矩阵D

输出: $\mathcal{X}$的k个类的划分

$S \leftarrow$ 均匀不放回的从$1, 2, ..., n$采样s个数字

$W \leftarrow D, K \leftarrow D^{-1}AD^{-1}$

$W_s \leftarrow W(S, S), K_s \leftarrow K(S, S)$

$\pi' \leftarrow$ 给定W_s和K_s在采样点上运行一个α近似算法

/* 把数据点靠到对应的中心点上 */

for $i = 1, ..., n$ **do**

  for $c = 1, ..., k$ **do**

$$d(a_i, m_c) \leftarrow K_{ii} - \frac{2\sum_{a_j \in \pi'_c} w_j K_{ij}}{\sum_{a_j \in \pi'_c} w_j} + \frac{\sum_{a_j, a_l \in \pi'_c} w_j w_l K_{jl}}{\left(\sum_{a_j \in \pi'_c} w_j\right)^2}$$

  end

  $j \leftarrow \operatorname*{argmin}_{c=1,2,...,k} d(a_i, m_c)$

  $Y_{ij} \leftarrow 1$

end

返回 Y

第三个贡献

- 对基于均匀采样和带权kernel k-means的谱聚类给出了更紧的理论界
- 给出了这一算法的MATLAB实现，并用实验验证了它的效率和解的质量

一个更紧的理论界

定理 (算法6的更紧理论界)

令$0 < \delta < 1/2$，$\alpha \geq 1$，$\beta > 0$是近似系数，假定我们用算法6采样s个点，且

$$s \geq \ln(\frac{1}{\delta})(1 + \frac{1}{n})/(\frac{\beta^2 m^2}{2\Delta^2\alpha^2} + \frac{\ln(1/\delta)}{n}) \tag{12}$$

则下式以至少$1 - 2\delta$的概率成立，

$$Ncut \leq (\alpha + \beta)Ncut^* + (\alpha + \beta - 1)c \tag{13}$$

其中$\Delta = \max_{i,j} \|\varphi(x_i) - \varphi(x_j)\|^2$是在映射空间中数据直径的平方，$m = \Psi_{OPT}(\mathcal{X})/n$是最优目标函数值的平均值，$c$是一个与划分无关的常数，$Ncut^*$是最优$Ncut$

目录

传统聚类

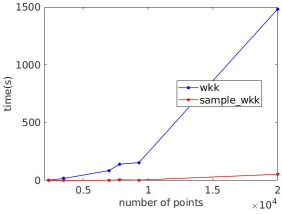
表 3: 数据量n，类数目k，维度d

数据集	名字	n	d	k
D_1	segment	2310	19	7
D_2	MnistData-05	3495	784	10
D_3	MnistData-10	6996	784	10
D_4	isolet5	7797	617	26
D_5	USPS	9298	256	10
D_6	letter-recognition	20000	16	26

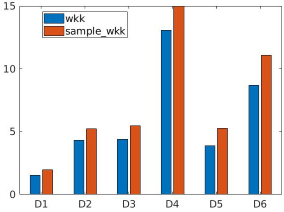
- 实验用于验证均匀采样的效率和聚类质量
- 算法：带权kernel k-means（均匀采样和不采样）
- 基于anchor的方法构建图
- 采样量：20%的数据点
- α近似算法：带权kernel k-means++
- 评价指标：时间(s)和Ncut

结果

1. 采样版本比不采样快了大概20∼50倍，而Ncut确并不大，不超过不采样的25%
2. 鉴于采样算法的高效性和合理的聚类质量，均匀采样是一个更好的选择



(A) 时间(s)



(B) Ncut

图 3: 谱聚类结果

目录

总结

1. (理论) 将经典的k-means++算法扩展到了带权重的情况，且给出了算法的理论证明
2. (理论) 在k-means问题上给出了更紧的理论界，并在数据有假设的情况下证明了这一算法的高效性
3. (理论) 将均匀采样的证明扩展到了谱聚类上，证明了更紧的理论界
4. (实验) 在k-means上对均匀采样、K-MC^2、Double-K-MC^2，和它们的kernel版给出了MATLAB实现，实验验证了这些算法的效率和有效性
5. (实验) 用MATLAB实现了基于带权kernel k-means的谱聚类算法和它对应的均匀采样版，实验说明了均匀采样版本的高效和合理的聚类质量

问题?

参考文献

D. Arthur and S. Vassilvitskii. k-means++: The advantages of careful seeding. In *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 1027–1035. Society for Industrial and Applied Mathematics, 2007.

O. Bachem, M. Lucic, S. H. Hassani, and A. Krause. Approximate k-means++ in sublinear time. 2016.

B. Bahmani, B. Moseley, A. Vattani, R. Kumar, and S. Vassilvitskii. Scalable k-means++. *Proceedings of the VLDB Endowment*, 5(7):622–633, 2012.

I. S. Dhillon, Y. Guan, and B. Kulis. *A unified view of kernel k-means, spectral clustering and graph cuts*. Citeseer, 2004.

S. Lloyd. Least squares quantization in pcm. *IEEE transactions on information theory*, 28(2):129–137, 1982.

M. Mohan and C. Monteleoni. Beyond the nystrom approximation: Speeding up spectral clustering using uniform sampling and weighted kernel k-means. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, IJCAI'17, pages 2494–2500. AAAI Press, 2017. ISBN 978-0-9992411-0-3. URL `http://dl.acm.org/citation.cfm?id=3172077.3172235`.

X. Y. Stella and J. Shi. Multiclass spectral clustering. In *null*, page 313. IEEE, 2003.