

# 大规模低质量多模态数据聚类

刘新旺

Email: xinwangliu@nudt.edu.cn

国防科技大学 计算机学院  
模式识别与机器智能教研室

2019年10月22日



# 报告内容

- 1 研究背景及现状
- 2 我们的工作
  - 缺失多核聚类算法 (AAAI-2017、IEEE TPAMI-2019)
  - 非完整多视图聚类 (IEEE TPAMI-2018、AAAI-19)
  - 基于 DNN 的近似大规模多核 K 均值聚类算法 (IJCAI-2017)
- 3 总结与展望

# 大数据特性

大数据具有如下特性：

- 数据量大：计算效率？
- 信息多：多源信息融合
- 数据质量低：数据缺失、噪声
- ...

聚类是大数据分析中的常用算法之一，聚焦于大规模低质量多模态数据的聚类算法研究。

# 聚类及其应用

## 聚类定义

聚类是将一组给定的数据依据它们的相似性划分为不同的簇，同一个簇中的样本差异性较小，不同簇中的样本差异性较大。

常见的聚类方法包括:

- K-均值聚类 / 核 K-均值聚类
- 谱聚类

这些聚类算法已成功应用于图像分割、异常检测、目标跟踪、场景发现、社交网络等诸多领域。

# 多视图表示

## 数据的多视图表示

影响聚类性能的关键是如何计算样本间的相似度，其依赖于**数据的特征表示**。为了克服某种类型特征的缺陷和不足，可以用来自不同类型的数据特征来描述样本。每个不同类型的特征被称为**视图**。

## 多视图表示的优越性

多视图聚类利用这些不同视图间的互补和兼容信息，充分发挥各自的优势，规避各自的局限，从而获得最优且稳健的聚类性能。

# 典型的多视图聚类算法

现有多视图聚类算法可以大致分为两类：

- 第一类算法通过低秩优化从多个视图中学习一个公共矩阵用于聚类 [1,2]。
- 第二类算法遵循多核学习框架，最优地组合这些视图用于聚类 [3,4,5,6]。

<sup>1</sup>Xia, Rongkai et al. “Robust multi-view spectral clustering via low-rank and sparse decomposition”.

In: AAAI. 2014.

<sup>2</sup>Zhou, Peng et al. “Recovery of Corrupted Multiple Kernels for Clustering”. In: IJCAI. 2015.

<sup>3</sup>Shi Yu et al. “Optimized Data Fusion for Kernel k-Means Clustering”. In: IEEE TPAMI (2012).

<sup>4</sup>Mehmet Gönen et al. “Localized Data Fusion for Kernel k-Means Clustering with Application to Cancer Biology”. In: NIPS. 2014.

<sup>5</sup>Xinwang Liu et al. “Multiple Kernel k-Means Clustering with Matrix-induced Regularization”. In: AAAI. 2016.

<sup>6</sup>Miaomiao Li, Xinwang Liu et al. “Multiple Kernel Clustering with Local Kernel Alignment Maximization”. In: IJCAI. 2016.

# 多核 $K$ -均值聚类 (Multiple Kernel k-means, MKKM)

多核  $K$ -均值聚类的优化目标为:

$$\min_{H \in \mathbb{R}^{n \times k}, \beta \in \mathbb{R}_+^m} \text{Tr} \left( K_\beta (I_n - HH^\top) \right) \quad \text{s.t.} \quad H^\top H = I_k, \beta^\top \mathbf{1}_m = 1, \quad (1)$$

其中  $K_\beta = \sum_{p=1}^m \beta_p^2 K_p$ ,  $\{K_p\}_{p=1}^m$  是  $m$  个预先指定的核矩阵,  $\beta = [\beta_1, \dots, \beta_m]^\top$  为每个基核的权重,  $H \in \mathbb{R}^{n \times k}$  是聚类指示矩阵。

求解问题 Eq.(1) 的交替优化框架

- i): 给定  $\beta$  优化  $H \iff$  特征值分解问题;
- ii): 给定  $H$ , 关于  $\beta$  的优化归结为如下具有线性约束的二次规划问题:

$$\min_{\beta \in \mathbb{R}_+^m} \sum_{p=1}^m \beta_p^2 \text{Tr} \left( K_p (I_n - HH^\top) \right) \quad \text{s.t.} \quad \beta^\top \mathbf{1}_m = 1. \quad (2)$$

# 报告内容

- 1 研究背景及现状
- 2 我们的工作
  - 缺失多核聚类算法 (AAAI-2017、IEEE TPAMI-2019)
  - 非完整多视图聚类 (IEEE TPAMI-2018、AAAI-19)
  - 基于 DNN 的近似大规模多核 K 均值聚类算法 (IJCAI-2017)
- 3 总结与展望

## 研究动机

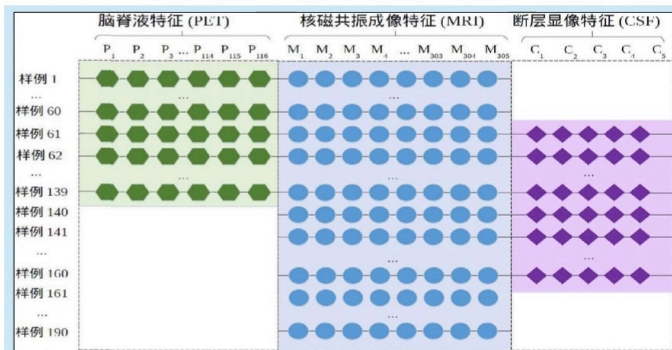


图1.阿尔兹魔症样例。每个样例由脑脊液特征、核磁共振成像特征及断层显像特征来描述。某些样例因病人没有做相应检查而缺失，对应于图中空白部分。

如何**最优地**组合这些**缺失视图**以获得更好的聚类结果？

# 缺失多核聚类算法—动机

## 现有算法

- ① 先对缺失部分进行填充，填充算法包括：零填充、均值填充、EM 填充，低秩填充等；
- ② 在填充后的数据上执行多核 K-均值算法；

缺陷：填充与聚类分离！

## 本文算法

提出面向聚类的填充算法以处理缺失多核聚类问题<sup>[a]</sup>。该算法：

- ① 利用聚类结果对每个核矩阵的缺失部分进行填充；
- ② 最优地组合填充后的核矩阵以产生聚类结果；

上述两个步骤交替地执行直至收敛。

<sup>a</sup>Xinwang Liu et al. “Multiple Kernel k-means with Incomplete Kernels”. In: AAAI. 2017.

# 缺失多核聚类算法—优化目标

缺失多核聚类算法的目标函数为:

$$\begin{aligned} \min_{\mathbf{H}, \boldsymbol{\beta}, \{K_p\}_{p=1}^m} & \text{Tr}(\mathbf{K}\boldsymbol{\beta}(\mathbf{I}_n - \mathbf{H}\mathbf{H}^\top)) \\ \text{s.t.} & \mathbf{H} \in \mathbb{R}^{n \times k}, \mathbf{H}^\top \mathbf{H} = \mathbf{I}_k, \boldsymbol{\beta}^\top \mathbf{1}_m = 1, \beta_p \geq 0, \\ & \boxed{K_p(s_p, s_p) = K_p^{(cc)}, K_p \succeq 0, \forall p} \end{aligned} \quad (3)$$

- 该算法将**聚类**和**填充**统一到同一目标函数。算法的终极目标是聚类，对核矩阵的缺失部分进行填充只是该算法的副产品。
- 聚类结果用于填充缺失核矩阵；填充后的核矩阵被最优地组合以产生更准确的聚类结果。这两个过程无缝对接并相互协调，以达到好的聚类结果。
- 该算法没有超参数需要调整，这使得它能够很好地适用于实际应用。

# 缺失多核聚类算法—求解算法

## 优化算法

提出一种三步交替优化算法来求解 Eq.(3) 中的优化问题，即：

- ① 固定  $\beta$  和  $\{K_p\}_{p=1}^m$ ，优化  $H$

$$\min_H \text{Tr}(K_\beta(I_n - HH^\top)) \quad \text{s.t.} \quad H \in \mathbb{R}^{n \times k}, H^\top H = I_k \quad (4)$$

- ② 固定  $\beta$  和  $H$ ，优化  $\{K_p\}_{p=1}^m$

$$\min_{K_p} \text{Tr}(K_p(I_n - HH^\top)) \quad \text{s.t.} \quad K_p(s_p, s_p) = K_p^{(cc)}, K_p \geq 0, \forall p. \quad (5)$$

- ③ 固定  $\{K_p\}_{p=1}^m$  和  $H$ ，优化  $\beta$

$$\min_\beta \sum_{p=1}^m \beta_p^2 \text{Tr}(K_p(I_n - HH^\top)) \quad \text{s.t.} \quad \beta^\top \mathbf{1}_m = 1, \beta_p \geq 0, \forall p. \quad (6)$$

# 缺失多核聚类算法—试验结果 (1/2)

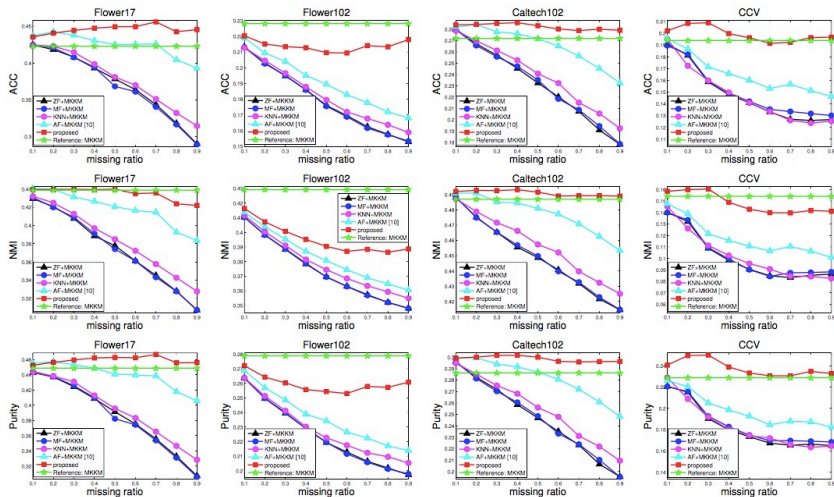


图: 各算法的聚类精度、互信息、纯度随缺失比率变化曲线。

# 缺失多核聚类算法—试验结果 (2/2)

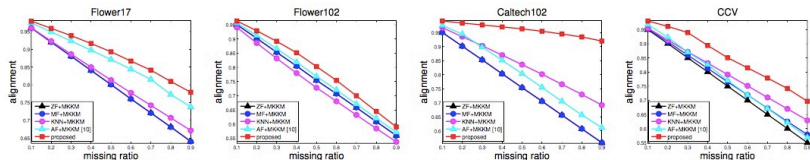


图: 各算法填充得到的核矩阵与真实矩阵间的对齐程度随缺失比率变化曲线

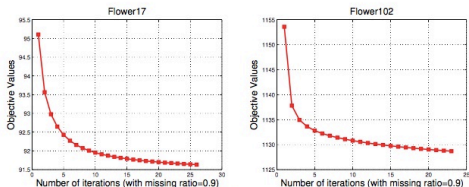


图: 本文提出算法的目标函数随迭代的变化曲线

本文全文及代码下载链接:

<http://www.esience.cn/people/liuxinwang/index.html>

# 非完整多视图聚类—研究动机

- 高额的计算和存储开销
- 过度复杂的填充模型
- 有限改进的聚类性能

首先定义第  $p$ -th ( $1 \leq p \leq m$ ) 个基聚类矩阵为

$$\mathbf{H}_p = [\mathbf{H}_p^{(o)\top}, \mathbf{H}_p^{(u)\top}]^\top \in \mathbb{R}^{n \times k}, \quad (1)$$

其中  $\mathbf{H}_p^{(o)} \in \mathbb{R}^{n_p \times k}$  可以对  $m$  个非完整核矩阵  $\{\mathbf{K}_p(\mathbf{s}_p, \mathbf{s}_p)\}_{p=1}^m$  执行核  $k$ -means 算法获取到,  $\mathbf{H}_p^{(u)} \in \mathbb{R}^{(n-n_p) \times k}$  表示  $\mathbf{H}_p$  的非完整部分, 它需要在在学的过程中填充。

## 非完整多视图聚类-算法模型

EE-IMVC 同时执行聚类和对  $\{\mathbf{H}_p^{(u)}\}_{p=1}^m$  的填充, 同时保持  $\{\mathbf{H}_p^{(o)}\}_{p=1}^m$  在学习的过程中保持不变。

$$\begin{aligned} & \max_{\mathbf{H}, \{\mathbf{W}_p, \mathbf{H}_p^{(u)}, \beta_p\}_{p=1}^m} \text{Tr} \left[ \mathbf{H}^\top \sum_{p=1}^m \beta_p \begin{pmatrix} \mathbf{H}_p^{(o)} \\ \mathbf{H}_p^{(u)} \end{pmatrix} \mathbf{W}_p \right] \\ & \text{s.t. } \mathbf{H} \in \mathbb{R}^{n \times k}, \mathbf{H}^\top \mathbf{H} = \mathbf{I}_k, \mathbf{W}_p \in \mathbb{R}^{k \times k}, \mathbf{W}_p^\top \mathbf{W}_p = \mathbf{I}_k, \\ & \mathbf{H}_p^{(u)} \in \mathbb{R}^{(n-n_p) \times k}, \mathbf{H}_p^{(u)\top} \mathbf{H}_p^{(u)} = \mathbf{I}_k, \boldsymbol{\beta} \in \mathbb{R}^m, \sum_{p=1}^m \beta_p^2 = 1, \beta_p \geq 0, \end{aligned} \quad (2)$$

其中  $\mathbf{H}$  和  $\mathbf{H}_p^{(u)}$  分别代表公共聚类矩阵和第  $p$  个基聚类矩阵的缺失部分,  $\mathbf{W}_p$  代表第  $p$  个置换矩阵, 用于最优地匹配  $\mathbf{H}_p$  和  $\mathbf{H}$ ,  $\boldsymbol{\beta} = [\beta_1, \dots, \beta_m]^\top$  是  $m$  个基聚类矩阵的权重。

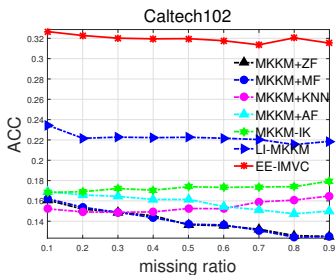
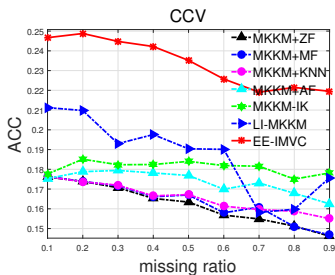
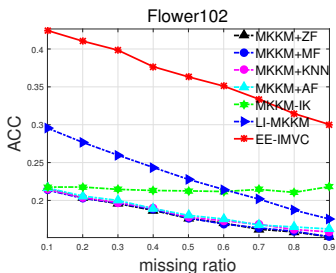
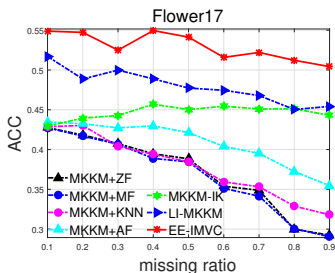
# 非完整多视图聚类—优化算法

设计了一个简单但计算有效的方法来解决该优化问题。

- 1) 固定  $\{\mathbf{W}_\rho, \mathbf{H}_\rho^{(u)}\}_{\rho=1}^m$  和  $\beta$  来优化  $\mathbf{H}$ ;
- 2) 固定  $\mathbf{H}$ ,  $\{\mathbf{H}_\rho^{(u)}\}_{\rho=1}^m$  和  $\beta$  来优化;
- 3) 固定  $\{\mathbf{W}_\rho\}_{\rho=1}^m$ ,  $\mathbf{H}$  和  $\beta$  来优化  $\{\mathbf{H}_\rho^{(u)}\}_{\rho=1}^m$ ;
- 4) 固定  $\mathbf{H}$  and  $\{\mathbf{W}_\rho, \mathbf{H}_\rho^{(u)}\}_{\rho=1}^m$  来优化  $\beta$ 。

算法从理论上保证具有（局部）最优解。

## 非完整多视图聚类—实验结果



# 基于 DNN 的近似大规模多核 K 均值聚类—动机

## 现有算法

- ① 多核聚类算法计算基核时的时间复杂度为  $\mathcal{O}(m * d^2 * n^2)$ ，其中  $m$ ， $d$  和  $n$  分别代表基核个数、样本维度和样本数量；
- ② 多核聚类算法需要将所有核加载进去进行优化学习，内存消耗为  $\mathcal{O}(m * n^2)$ ；
- ③ 多核聚类算法中每次迭代需要对组合核进行奇异值分解产生指示矩阵，时间复杂度达到  $\mathcal{O}(n^3)$ ；

内存和时间的巨大开销使得多核聚类算法无法应用于大规模任务！

## 本文贡献

使用深度神经网络模拟多核聚类中生成核和聚类的过程 [a]，从而减小时间和内存开销。

<sup>a</sup>Yueqing Wang, Xinwang Liu et al. “Approximate Large-scale Multiple Kernel k-means using Deep Neuron Network”. In: IJCAI. 2017.

# 基于 DNN 的近似大规模多核 K 均值聚类—示意图

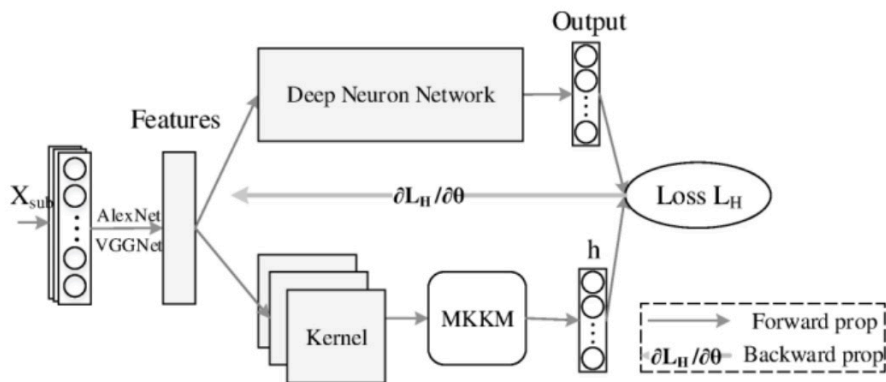


图 5.1 近似多核聚类算法架构

# 基于 DNN 的近似大规模多核 K 均值聚类算法

---

## 算法 5.1 训练阶段算法.

---

**Input:**  $\mathbf{X}_{sub}$

**Output:**  $\theta$ ,  $\mathbf{W}_{AE}$  and  $\mathbf{b}_{AE}$

- 1: 通过子集  $\mathbf{X}_{sub}$  产生  $m$  个基核  $\{\mathbf{K}_i\}_{i=1}^m$ ;
  - 2: 使用 MKKM 算法计算子集对应的指示矩阵  $\mathbf{H}_{sub}$ ;
  - 3: 随机初始化网络参数  $\theta$ ;
  - 4: **while** 未收敛 **do**
  - 5:     正向计算网络输出  $f_{\theta}(\mathbf{X}_{sub})$ ;
  - 6:     反向传播计算  $\partial J/\partial \theta$ ;
  - 7:     更新  $\theta$ ;
  - 8: **end while**
  - 9: 使用  $f_{\theta}(\mathbf{X}_{sub})$  作为输入训练自编码器, 得到自编码网络的参数  $\mathbf{W}_{AE}$  和  $\mathbf{b}_{AE}$ .
- 

---

## 算法 5.2 测试阶段算法 (Test stage)

---

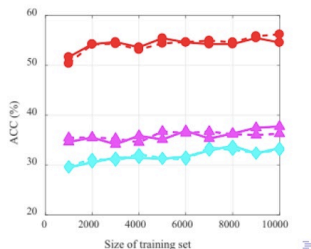
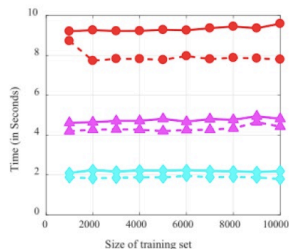
**Input:**  $\mathbf{X}$ ,  $\theta$ ,  $\mathbf{W}_{AE}$  和  $\mathbf{b}_{AE}$

**Output:**  $\theta$

- 1: 计算  $\mathbf{H}_o = f_{\theta}(\mathbf{X})$ ;
  - 2: 对  $\mathbf{H}_o$  使用 PCA 或者自编码器进行降维, 得到  $\tilde{\mathbf{H}}$ ;
  - 3: 对  $\tilde{\mathbf{H}}$  使用  $k$ -means 聚类得到结果.
-

# 基于 DNN 的近似大规模多核 K 均值聚类—试验结果

	ACC (%)					
	mnist	cifar100	102flowers	birds200	caltech256	ImageNet
$k$ -means	64.48	31.32	50.71	29.54	48.63	36.28
PCA+ $k$ -means	54.38	28.82	48.61	25.15	44.52	34.80
LSC- $k$	<b>78.69</b>	30.54	49.13	24.52	41.28	30.37
RMKMC	-	-	36.05	20.80	41.50	-
AMGL	-	-	53.65	22.32	41.76	-
MLAN	-	-	52.50	30.81	36.98	-
CKM	60.57	30.14	49.28	27.56	48.02	33.21
MKKM	-	-	51.72	29.37	-	-
Ours	70.22	<b>37.37</b>	55.79	<b>31.34</b>	54.30	38.84
Ours+PCA	70.18	37.15	<b>55.99</b>	31.33	<b>54.45</b>	<b>38.92</b>



# 报告内容

## 1 研究背景及现状

## 2 我们的工作

- 缺失多核聚类算法 (AAAI-2017、IEEE TPAMI-2019)
- 非完整多视图聚类 (IEEE TPAMI-2018、AAAI-19)
- 基于 DNN 的近似大规模多核 K 均值聚类算法 (IJCAI-2017)

## 3 总结与展望

## 值得探索的方向

- 深度嵌入聚类、深度单分类
- 基于深度神经网络的迁移学习
- 深度核学习
- 深度学习的泛化性能分析（深度：表示能力更强、信息损失更重）

谢谢! 请多批评指正

xinwangliu@nudt.edu.cn

## 部分参考文献

- 1 **Xinwang Liu** et. al.: Late Fusion Incomplete Multi-view Clustering. [IEEE TPAMI 2018](#). (CCF Rank A)
- 2 **Xinwang Liu** et. al.: Multiple Kernel  $k$ -means with Incomplete Kernels. [IEEE TPAMI 2019](#). (CCF Rank A)
- 3 **Xinwang Liu** et. al.: Absent Multiple Kernel Learning Algorithms. [IEEE TPAMI 2019](#). (CCF Rank A)
- 4 **Xinwang Liu** et. al.: Efficient and Effective Incomplete Multi-view Clustering. [AAAI2019](#). (CCF Rank A)
- 5 **Xinwang Liu** et. al.: Multiple Kernel  $k$ -means with Incomplete Kernels. [AAAI2017](#). (CCF Rank A)
- 6 **Xinwang Liu** et. al.: Optimal Neighborhood Kernel Clustering with Multiple Kernels. [AAAI2017](#). (CCF Rank A)
- 7 **Xinwang Liu** et. al.: Efficient and Effective Regularized Incomplete Multi-view Clustering. [IEEE TPAMI 2019](#). (Major Revision) (CCF Rank A)