

多视图非线性聚类及其应用

王昌栋

中山大学 数据科学与计算机学院

changdongwang@hotmail.com

<http://sdcs.sysu.edu.cn/node/2465>

March 8, 2017

目录

- 研究背景及研究意义

- 非线性聚类研究

Multi-Exemplar Affinity Propagation, TPAMI 2013.

- 多视图聚类研究

Multi-View Clustering Based on Belief Propagation, TKDE 2016.

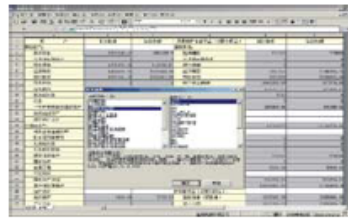
- 多视图非线性聚类

- 应用案例

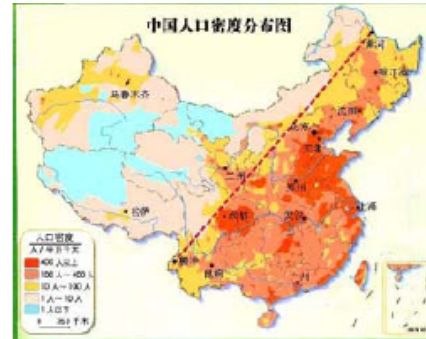
信息大爆炸时代下的数据处理



银行数据



企业报表



人口数据



如何有效地
分析和利用
这些数据？



信息大爆炸



交通数据



股票行情



多媒体数据

将所获得的数据分类是一种有效简化数据处理和分析的方法

数据聚类

- 数据聚类是模式识别中最基本的分类方法之一

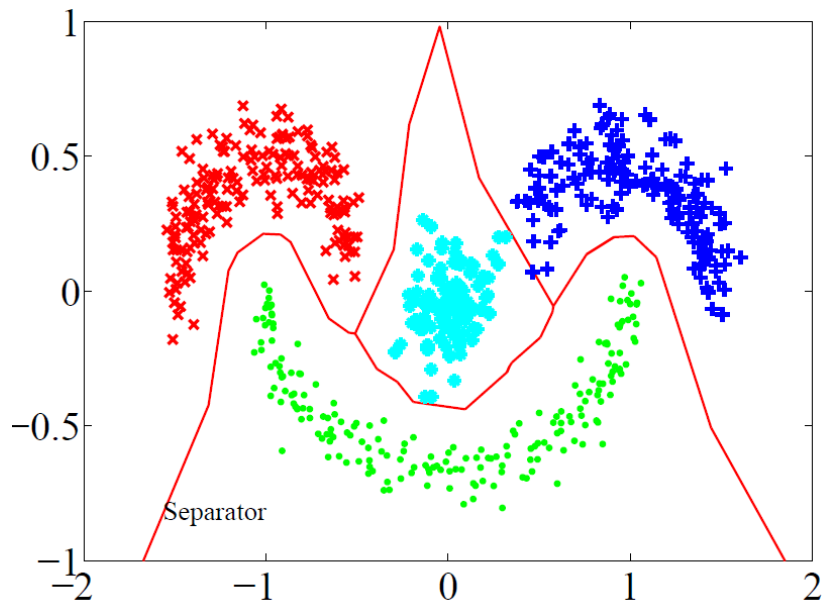


加菲猫 &
史努比

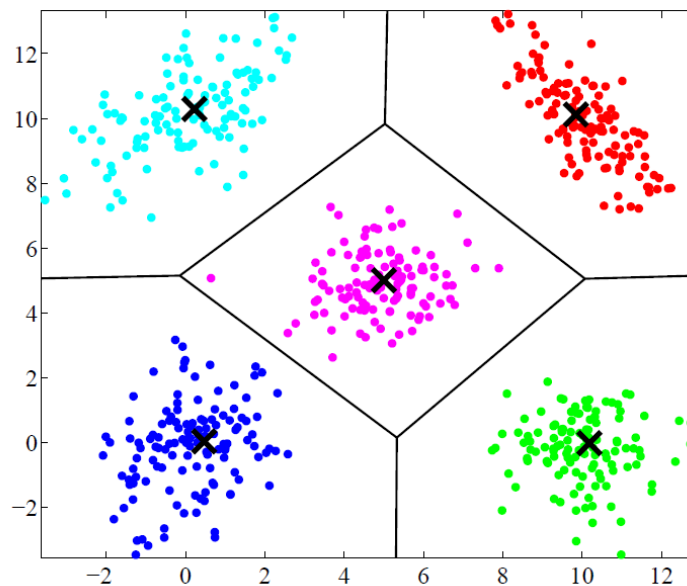
在没有先验知识监督的情况下，聚类算法将上述图像自动分成两类



非线性聚类定义



(a) 非线性可分数据集



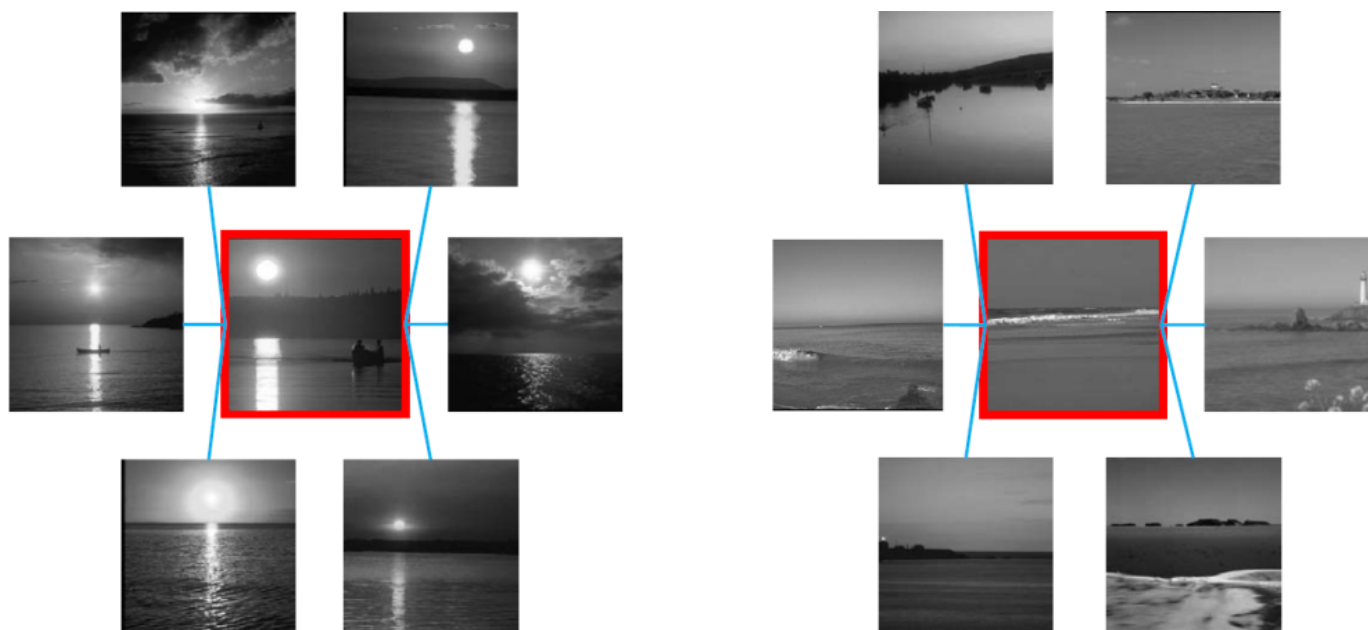
(b) 线性可分数据集

定义：对于一个数据集，若至少包含一个**非凸形状边界**的类，则称该数据集为**非线性可分**的。

线性可分的数据集可以通过**一个中心点表达一个类**，
但是非线性可分的数据集**无法通过一个中心点表达一个类**。

非线性聚类研究意义

- 由于现实数据的复杂流形分布，非线性聚类是最流行且最被广泛研究的聚类问题之一

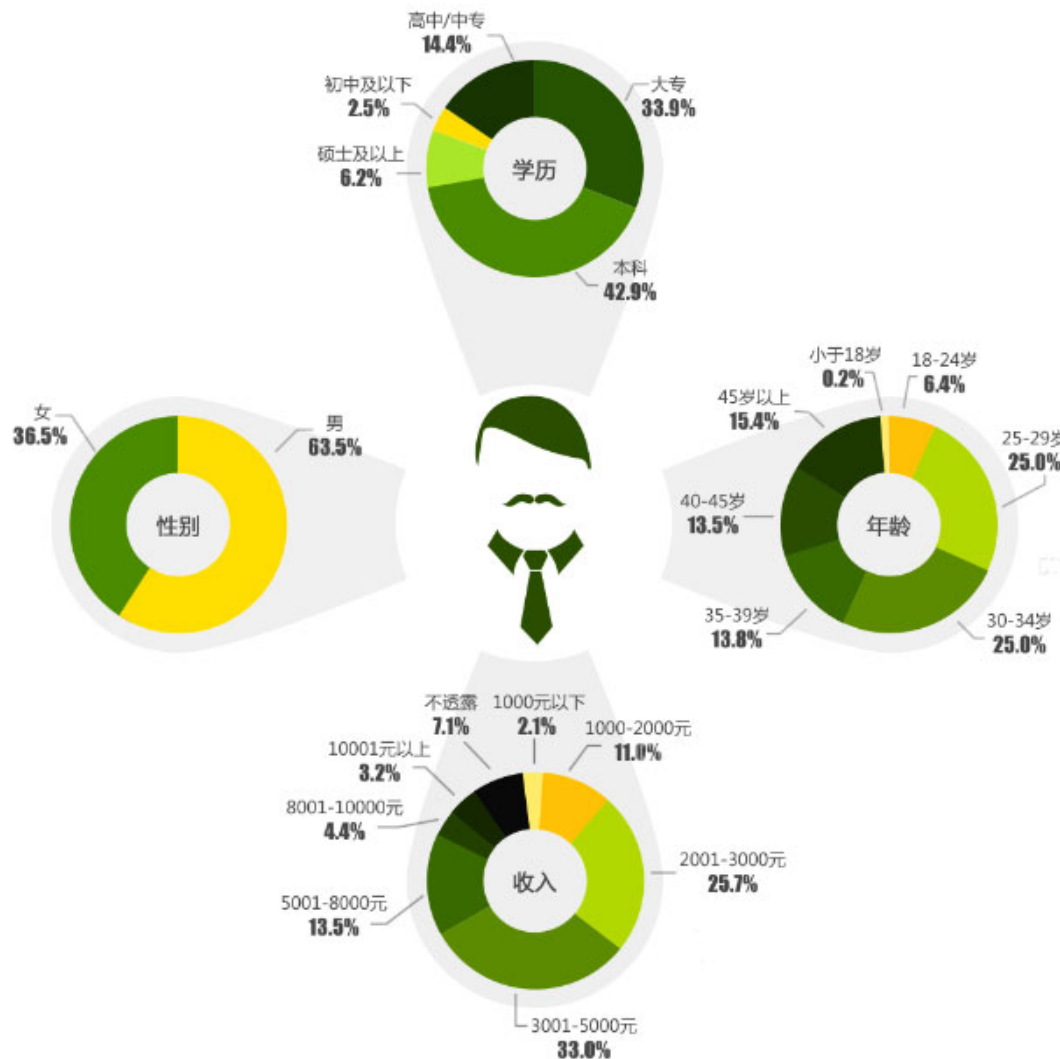


海岸场景的两个典型主题：含有及不含有日落/日出。
无法通过线性可分的单中心点模型进行聚类



非线性聚类的例子

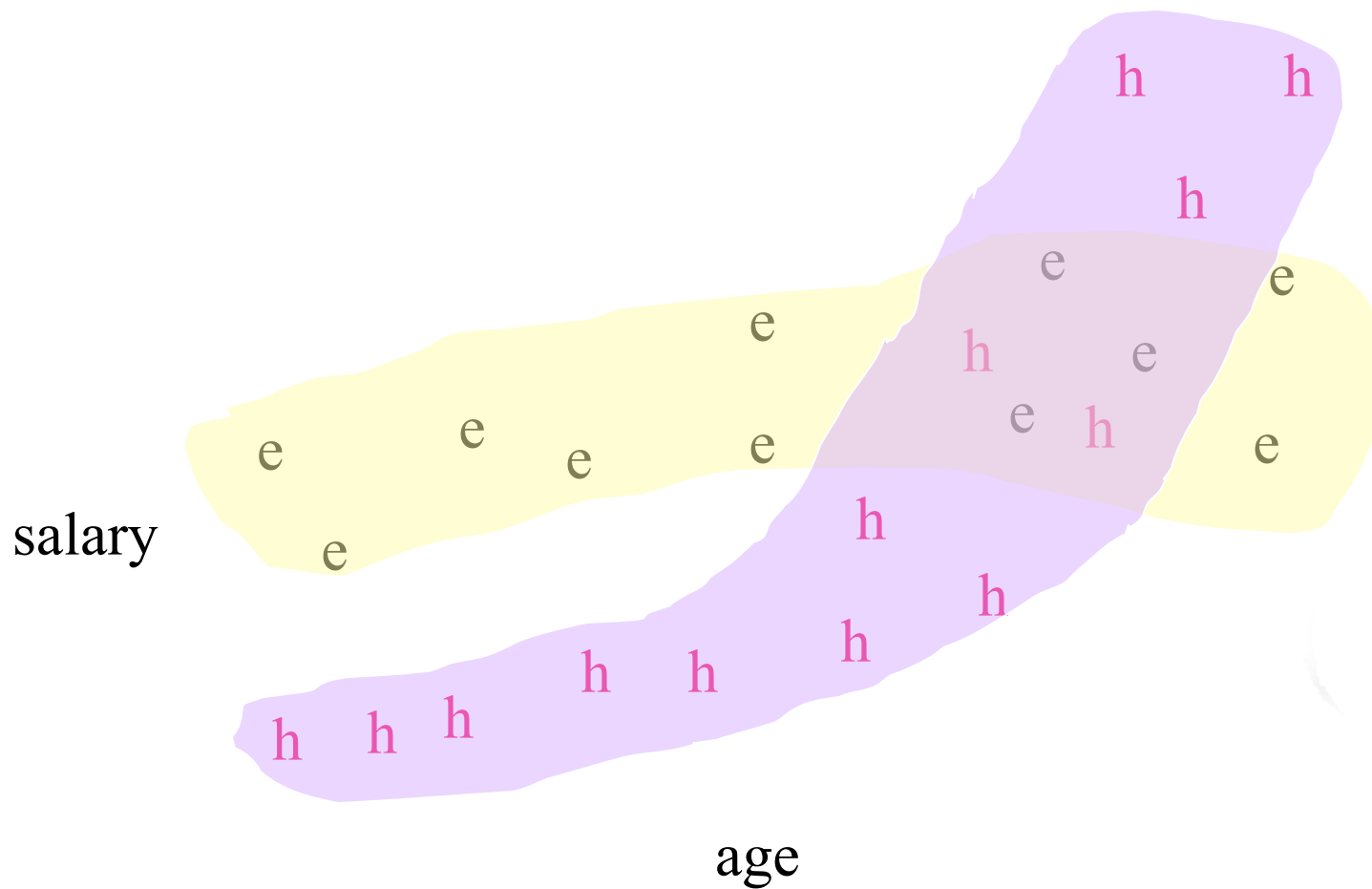
● 用户群体特性分析：



非线性聚类的例子: Stanford Salaries

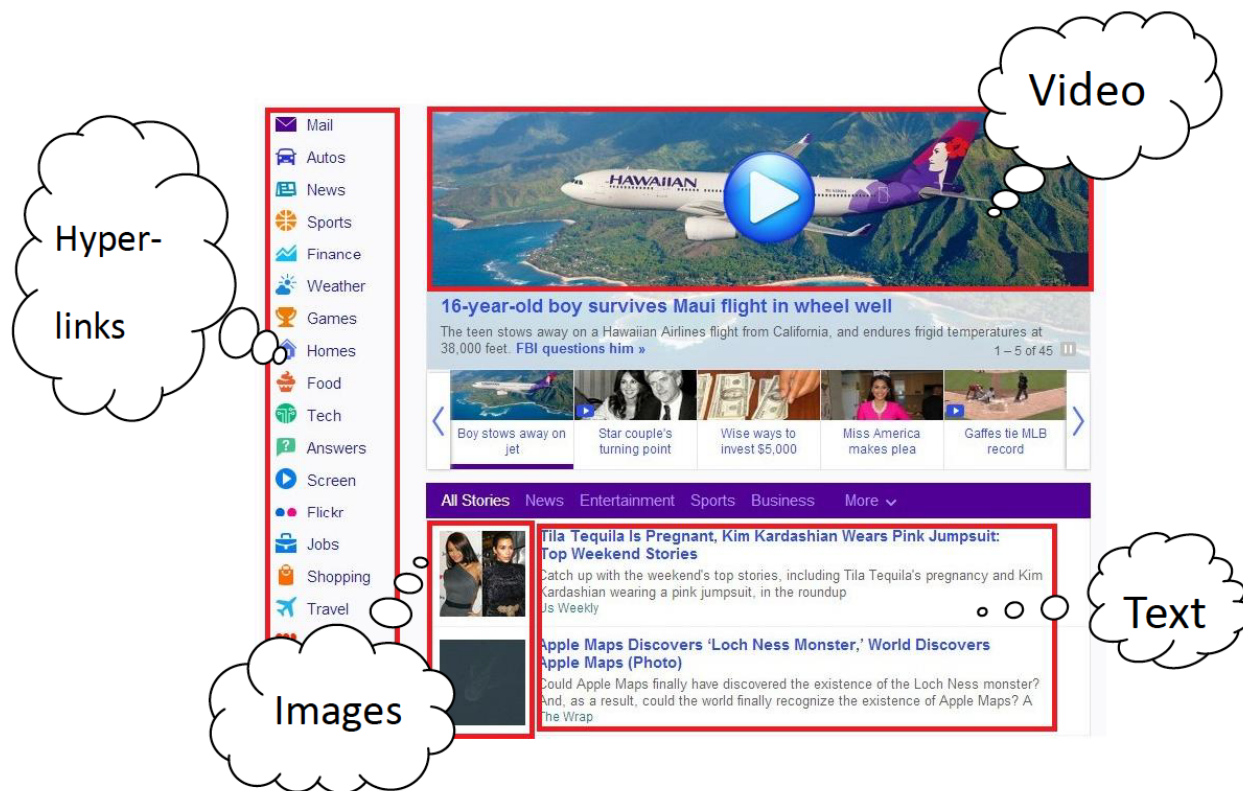


中山大學
SUN YAT-SEN UNIVERSITY



From <http://www.mmnds.org>

多视图数据聚类



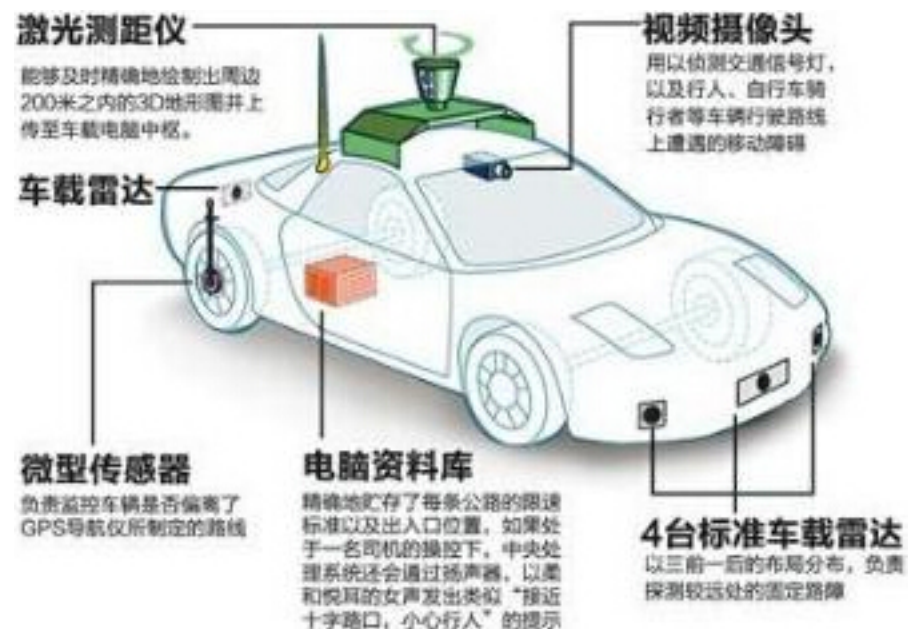
多视图数据：网页数据通常同时包含**超链接**、**视频**、**图像**、**文本**等元素，每一种元素便是该网页的一个**视图**。不同视图之间存在着很强的**互补性**。

多视图聚类：充分融合不同视图的信息进行聚类。

在大数据时代，由于大数据的**Variety**（**多样性**），多视图数据分析，特别是无监督聚类，是另一个重要的研究课题。

多视图数据聚类的例子

● 无人驾驶技术:



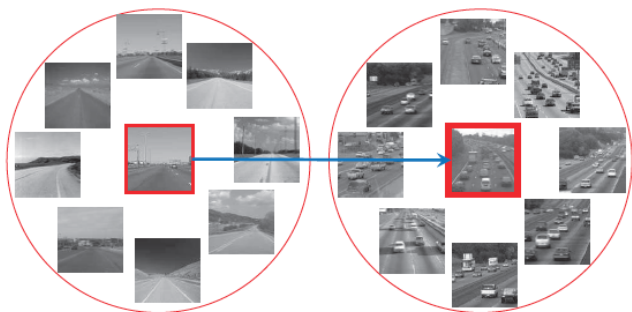
多中心点近邻传播算法 (MEAP)



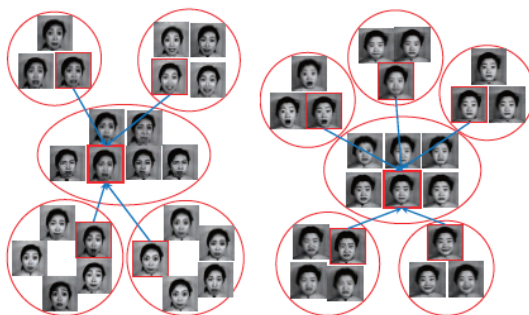
中山大學
SUN YAT-SEN UNIVERSITY

- 若数据集中的某些类**含有多个子类**，则称该数据集具有**多子类结构**。含有多个子类的类显然无法通过一个中心点来表达。从而，由非线性可分的定义，**具有多子类结构的数据集也是非线性可分的**。
- 具有多子类结构的例子

在自然场景图像



人脸图像



手写数字



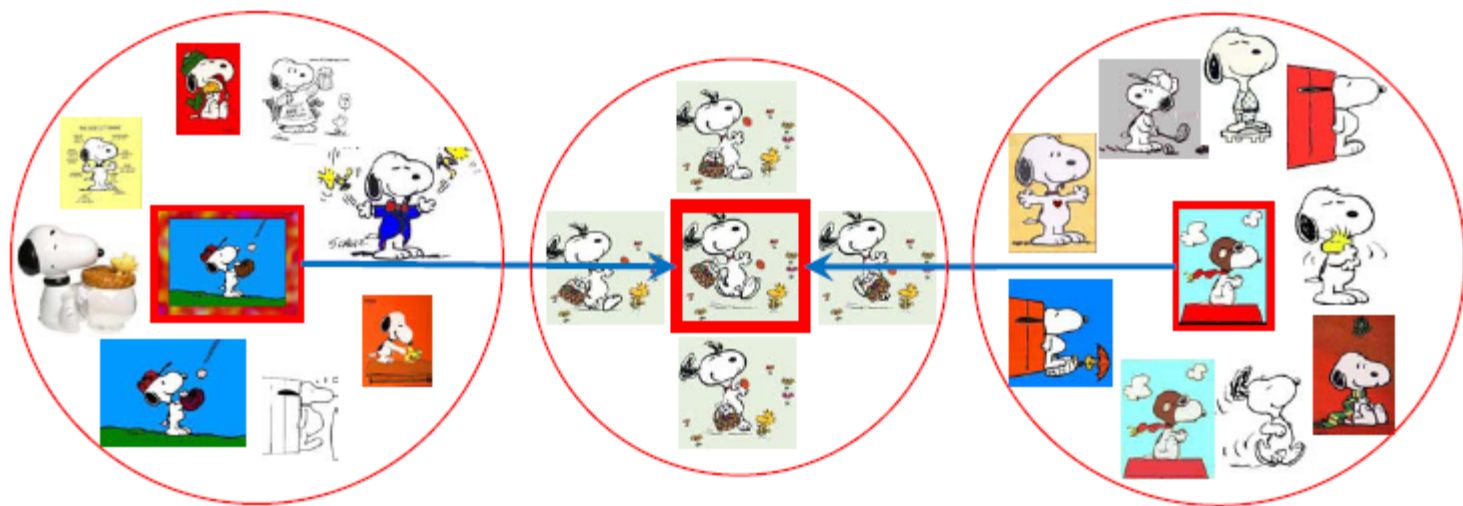
多中心点近邻传播算法 (MEAP)



中山大學
SUN YAT-SEN UNIVERSITY

- 刻画多子类结构的一个有效方法就是多中心点模型：

一个中心点表达一个对应的子类，再通过合并多个中心点来表达一个类



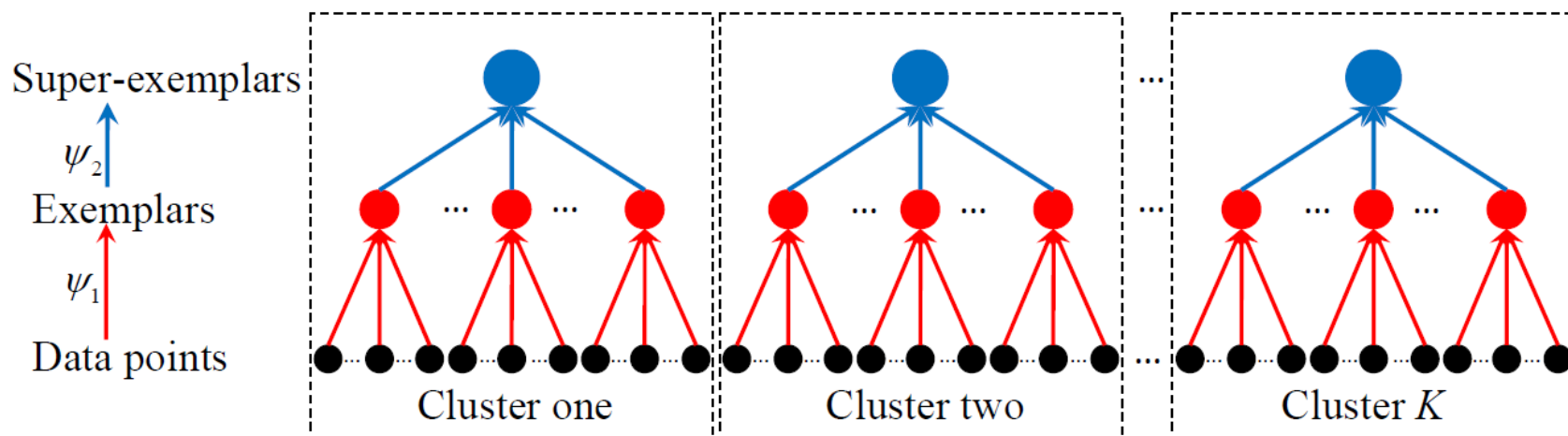
多中心点模型对史努比多子类结构的建模：三个中心点对应三个子类

多中心点近邻传播算法 (MEAP)



中山大學
SUN YAT-SEN UNIVERSITY

- 我们提出一个样本点->中心点->超级中心点的双层结构的多中心点模型



多中心点双层结构：每个样本点分配给最合适的中心点，同时每个中心点分配给最合适的超级中心点；每个类都由自动确定数量的中心点和一个超级中心点来建模

多中心点近邻传播算法 (MEAP)



中山大学
SUN YAT-SEN UNIVERSITY

多中心点模型的优化目标:

$$\mathcal{S}(C) = \mathcal{S}_1 + \mathcal{S}_2 + \text{three constraints}$$

$$= \sum_{i=1}^N \sum_{j=1}^N S_{ij}(c_{ij}) + \sum_{i=1}^N I_i(c_{i1}, \dots, c_{iN})$$

$$+ \sum_{j=1}^N E_j(c_{1j}, \dots, c_{Nj}) + \sum_{k=1}^N F_k(c_{11}, \dots, c_{NN}).$$

$$S_{ij}(c_{ij}) = \begin{cases} s_{ij}, & \text{if } i \neq j \text{ \& } c_{ij} \neq 0, \\ s_{ii} + l_{ic_{ii}}, & \text{if } i = j \text{ \& } c_{ii} \neq 0, \\ 0, & \text{otherwise.} \end{cases}$$

$$c_{ij} = \begin{cases} 1, & \text{if } j \text{ is an exemplar of } i \quad \forall i \neq j, \\ 0, & \text{otherwise,} \end{cases}$$

$$c_{ii} = \begin{cases} k \in \{1, \dots, N\}, & \text{if } k \text{ is a superexemplar of } i, \\ 0, & \text{if } i \text{ is not an exemplar.} \end{cases}$$

中心点N选1约束

$$I_i(c_{i1}, \dots, c_{iN}) = \begin{cases} -\infty & \text{if } \sum_{j=1}^N [c_{ij} \neq 0] \neq 1 \\ 0 & \text{otherwise.} \end{cases}$$

中心点一致性约束

$$E_j(c_{1j}, \dots, c_{Nj}) = \begin{cases} -\infty & \text{if } c_{jj} = 0 \text{ but } \exists i : c_{ij} = 1 \\ 0 & \text{otherwise.} \end{cases}$$

超级中心点一致性约束

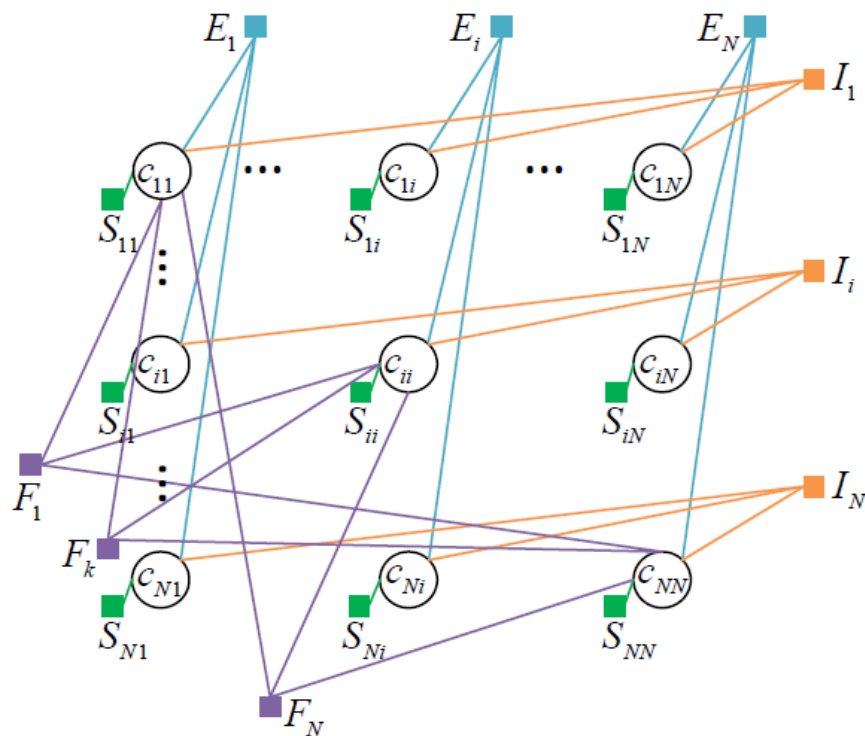
$$F_k(c_{11}, \dots, c_{NN}) = \begin{cases} -\infty & \text{if } c_{kk} \neq k \text{ but } \exists i : c_{ii} = k \\ 0 & \text{otherwise.} \end{cases}$$

多中心点近邻传播算法 (MEAP)



中山大学
SUN YAT-SEN UNIVERSITY

- 我们采用因子图 (Factor graph) 表达该多中心点模型



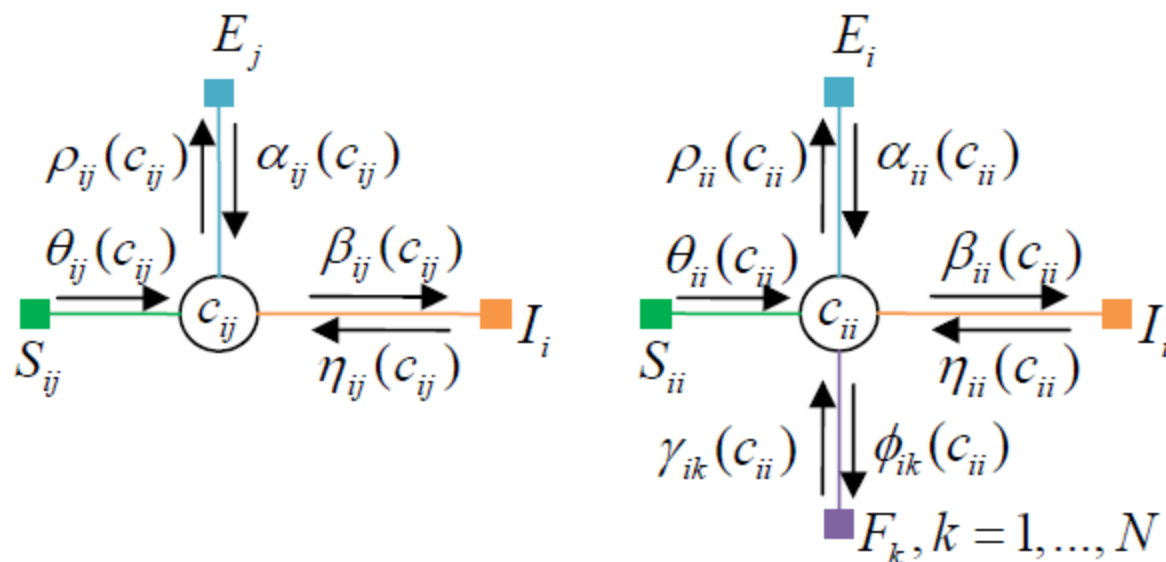
采用因子图表达多中心点双层结构：圆形的变量结点表示中心点/超级中心点分配情况，而方形的函数结点表示样本点 $\rightarrow$ 中心点 $\rightarrow$ 超级中心点的双层约束结构

多中心点近邻传播算法 (MEAP)



中山大学
SUN YAT-SEN UNIVERSITY

- 我们采用了最大和置信传播算法来解多中心点模型的最优化问题



MEAP算法的消息传递：通过不断地在变量结点与函数结点之间传递消息，使得多中心点模型最优化，从而得到最优的样本点->中心点->超级中心点分配函数

多中心点近邻传播算法 (MEAP)

● 最大和置信传播算法

消息传递：变量节点->函数节点

$$\mu_{x \rightarrow f}(x) \leftarrow \sum_{h \in \text{ne}(x) \setminus \{f\}} \mu_{h \rightarrow x}(x)$$

即除了该函数之外所有其它近邻函数发送到该变量的消息值之和。

消息传递：函数节点->变量节点

$$\mu_{f \rightarrow x}(x) \leftarrow \max_{X \setminus \{x\}} \left[f(X) + \sum_{y \in X \setminus \{x\}} \mu_{y \rightarrow f}(y) \right]$$

涉及除了该自变量之外所有其它自变量的最大化操作。



多中心点近邻传播算法 (MEAP)

$$i \neq j \quad \tilde{\rho}_{ij} \leftarrow s_{ij} - \max \left[\max_{j' \notin \{j,i\}} [s_{ij'} + \tilde{\alpha}_{ij'}], \max_{m \in \{1, \dots, N\}} [l_{im} + \tilde{\gamma}_{im}] + s_{ii} + \tilde{\alpha}_{ii} \right]$$

$$\forall i = 1, \dots, N, \quad \tilde{\rho}_i^k \leftarrow s_{ii} + l_{ik} - \max_{i' \neq i} [s_{ii'} + \tilde{\alpha}_{ii'}] + \tilde{\gamma}_{ik}$$

$$k = 1, \dots, N \quad \tilde{\alpha}_{ii} \leftarrow \sum_{i' \neq i} \max [0, \tilde{\rho}_{i'i}]$$

$$\tilde{\phi}_{ik} \leftarrow \min \left[l_{ik} - \max_{m \neq k} [l_{im} + \tilde{\gamma}_{im}], \tilde{\alpha}_{ii} + \tilde{\rho}_i^k - \tilde{\gamma}_{ik} \right]$$

$$\tilde{\gamma}_{kk} \leftarrow \sum_{i' \neq i} \max [0, \tilde{\phi}_{i'k}]$$

$$\tilde{\gamma}_{ik} \leftarrow \min \left[0, \tilde{\phi}_{kk} + \sum_{i' \notin \{i,k\}} \max [0, \tilde{\phi}_{i'k}] \right], k \neq i,$$

$$\begin{aligned} \hat{c}_{ij} &= \arg \max_{c_{ij}} [\theta_{ij}(c_{ij}) + \alpha_{ij}(c_{ij}) + \eta_{ij}(c_{ij})] \\ &= \begin{cases} 1 & \text{if } \tilde{\alpha}_{ij} + \tilde{\rho}_{ij} \geq 0 \\ 0 & \text{otherwise} \end{cases} \quad \forall i \neq j, \end{aligned} \quad (32)$$

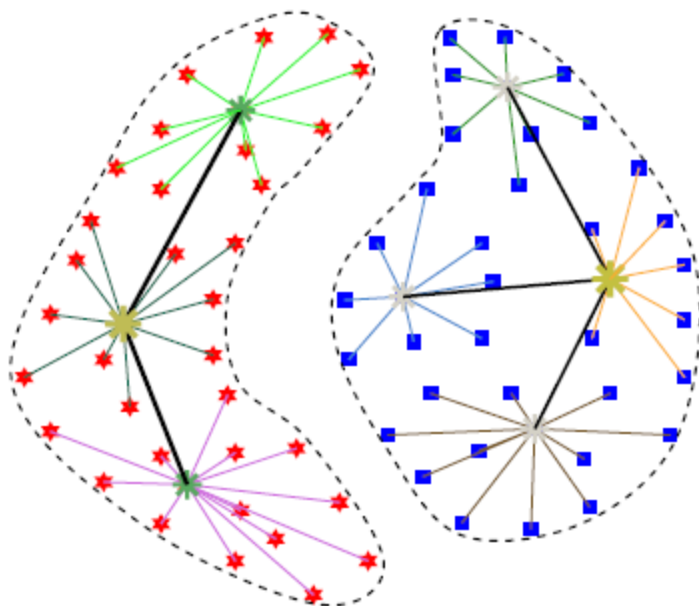
$$\begin{aligned} \hat{c}_{ii} &= \arg \max_{c_{ii}} \left[\theta_{ii}(c_{ii}) + \alpha_{ii}(c_{ii}) + \eta_{ii}(c_{ii}) + \sum_{k=1}^N \gamma_{ik}(c_{ii}) \right] \\ &= \begin{cases} \arg \max_k \tilde{\rho}_i^k & \text{if } \tilde{\alpha}_{ii} + \max_k \tilde{\rho}_i^k \geq 0 \\ 0 & \text{otherwise.} \end{cases} \end{aligned} \quad (33)$$

通过采用参数最大化的形式，取能够使得传递于每一个变量节点的消息之和最大化的变量取值，从而得到最终的类标分配向量。

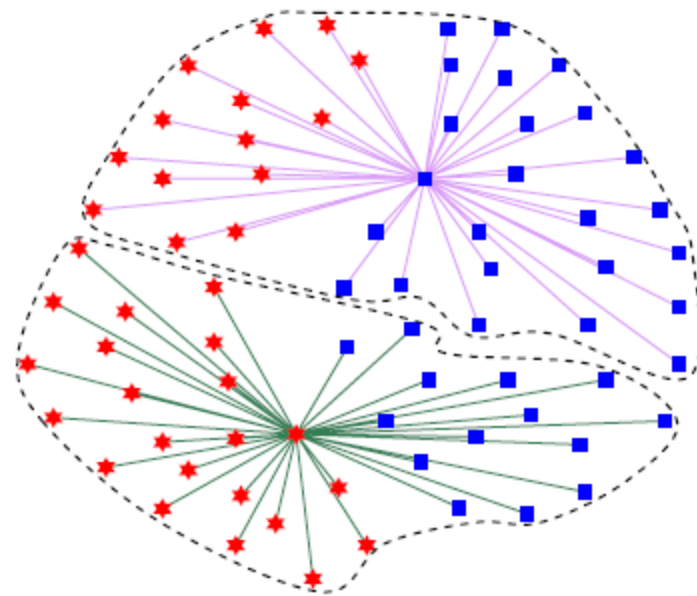
多中心点近邻传播算法 (MEAP)



中山大學
SUN YAT-SEN UNIVERSITY



(a) MEAP算法



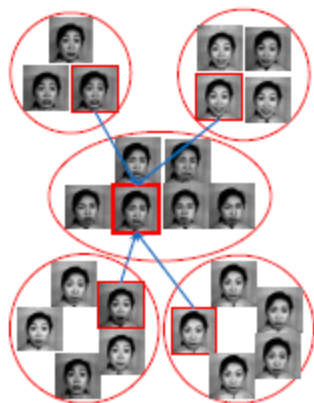
(b) AP算法

MEAP算法对比AP算法：通过多中心点模型，MEAP算法可以得到更优的划分边界，也即类边缘最大化

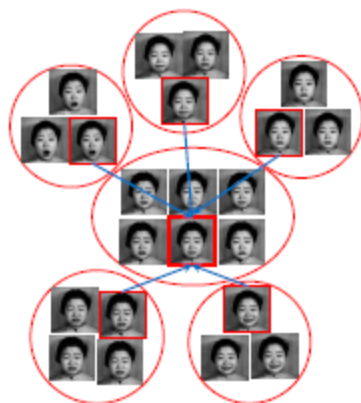
多中心点近邻传播算法 (MEAP)



中山大学
SUN YAT-SEN UNIVERSITY



(a) KA类



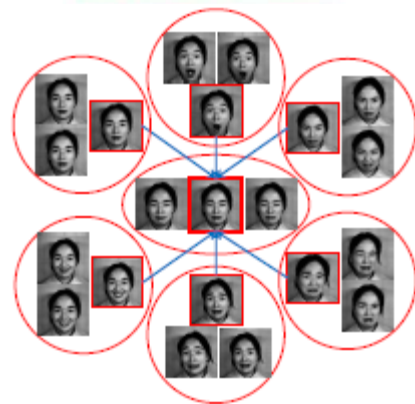
(b) KL类



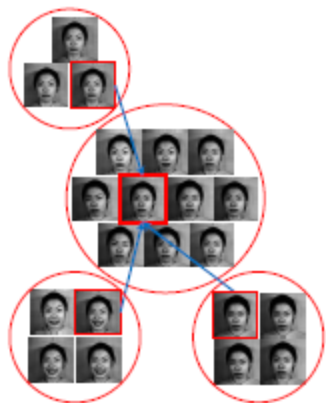
(c) KM类



(d) KR类



(e) MK类



(f) NA类



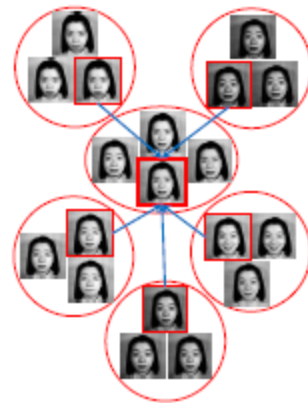
(g) NM类



(h) TM类



(i) UY类



(j) YM类

MEAP算法在JAFPE人脸库上的聚类结果：多中心点模型有效地刻画了因不同表情产生的多子类结构



多视图相似性传播 (MVAP)

● 多视图聚类方法

1. 相似性矩阵串联
2. 视图约束策略
3. 张量分解
4. 多视图谱聚类

Clustering quality
within individual
views

Balance?

Clustering consistency
across different views

多视图相似性传播 (MVAP)

Definition

Given m similarity matrixes $\tilde{S}^p \in \mathbb{R}^{n \times n}$, the goal of multi-view clustering can be transferred to find m cluster assignment vectors $\{\mathbf{c}^1, \mathbf{c}^2, \dots, \mathbf{c}^m\}$.

The p -th cluster assignment vector $\mathbf{c}^p = [\mathbf{c}_1^p, \mathbf{c}_2^p, \dots, \mathbf{c}_n^p]$ assigns each data point $\mathbf{x}_i$ to the most suitable exemplar (i.e. the $\mathbf{c}_i^p$ -th data point) in the p -th view, and these cluster assignment vectors should be as consistent as possible.

我们提出一个多视图相似性传播(Multi-View Affinity Propagation, MVAP)算法解决上述问题。



多视图相似性传播 (MVAP)

● 模型的目标函数

视图内聚
类质量

视图间聚
类一致性

$$J(\{\mathbf{c}^1, \dots, \mathbf{c}^m\}) = \sigma \sum_{p=1}^m J_w(\mathbf{c}^p) + (1 - \sigma) \sum_{p=1, q=1, p \neq q}^m J_b(\mathbf{c}^p, \mathbf{c}^q)$$

$$J_w(\mathbf{c}^p) = \sum_{i=1}^n \tilde{S}^p(i, \mathbf{c}_i^p) + \sum_{k=1}^n \tilde{\delta}_k^p(\mathbf{c}^p)$$

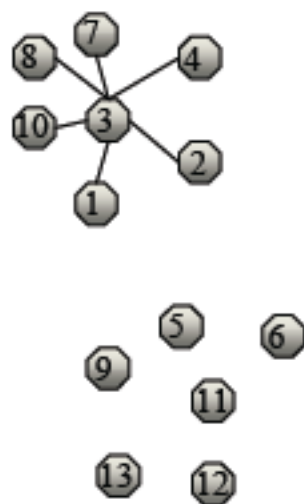
$$J_b(\mathbf{c}^p, \mathbf{c}^q) = \sum_{i=1}^n \tilde{L}^{pq}(\mathbf{c}_i^p, \mathbf{c}_i^q)$$

$$\tilde{\delta}_k^p(\mathbf{c}^p) = \begin{cases} -\infty & \text{If } \mathbf{c}_k^p \neq k \text{ but } \exists i : \mathbf{c}_i^p = k \\ 0 & \text{Otherwise} \end{cases}$$

$$\begin{aligned} \tilde{L}^{pq}(i, j) &= \text{JaccardSim}(\mathcal{N}_k^p(\mathbf{x}_i), \mathcal{N}_k^q(\mathbf{x}_j)) \\ &= \frac{|\mathcal{N}_k^p(\mathbf{x}_i) \cap \mathcal{N}_k^q(\mathbf{x}_j)|}{|\mathcal{N}_k^p(\mathbf{x}_i) \cup \mathcal{N}_k^q(\mathbf{x}_j)|} \end{aligned}$$

$$J(\{\mathbf{c}^1, \dots, \mathbf{c}^m\}) = \sum_{p=1}^m \left(\sum_{i=1}^n \sigma \cdot \tilde{S}^p(i, \mathbf{c}_i^p) + \sum_{k=1}^n \sigma \cdot \tilde{\delta}_k^p(\mathbf{c}^p) \right) + \sum_{p=1, q=1, p \neq q}^m \left(\sum_{i=1}^n (1 - \sigma) \cdot \tilde{L}^{pq}(\mathbf{c}_i^p, \mathbf{c}_i^q) \right)$$

多视图相似性传播 (MVAP)

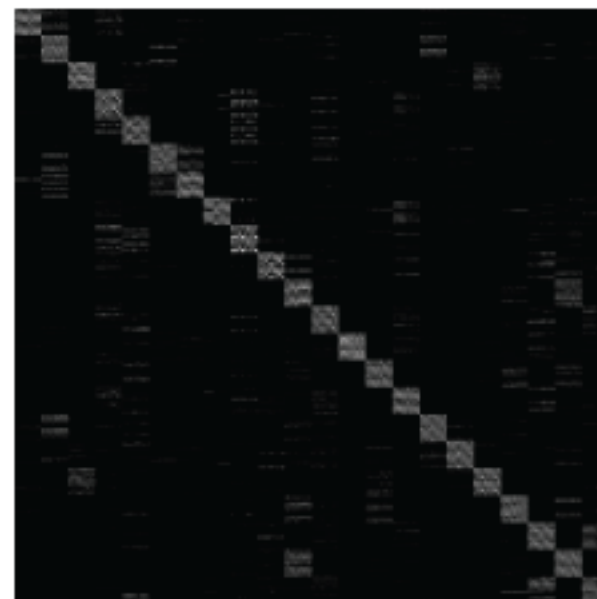


The p -th view



The q -th view

(a) $\tilde{L}^{pq}(3, 3)$.



(b) VIS/NIR dataset.

Figure : Illustration of data view consistency: (a) Data view consistency across the p -th and q -th data views:

$\tilde{L}^{pq}(3, 3) = \frac{|\{1, 2, 3, 4, 7, 8, 10\} \cap \{2, 3, 4, 5, 7, 8, 10\}|}{|\{1, 2, 3, 4, 7, 8, 10\} \cup \{2, 3, 4, 5, 7, 8, 10\}|} = \frac{3}{4}$; (b) Data consistency matrix of a VIS/NIR dataset consisting of the VIS and NIR face images.



多视图相似性传播 (MVAP)

$$J(\{\mathbf{c}^1, \dots, \mathbf{c}^m\}) = \sum_{p=1}^m \left(\sum_{i=1}^n S^p(i, \mathbf{c}_i^p) + \sum_{k=1}^n \delta_k^p(\mathbf{c}^p) \right) + \sum_{p=1, q=1, p \neq q}^m \left(\sum_{i=1}^n L^{pq}(\mathbf{c}_i^p, \mathbf{c}_i^q) \right)$$

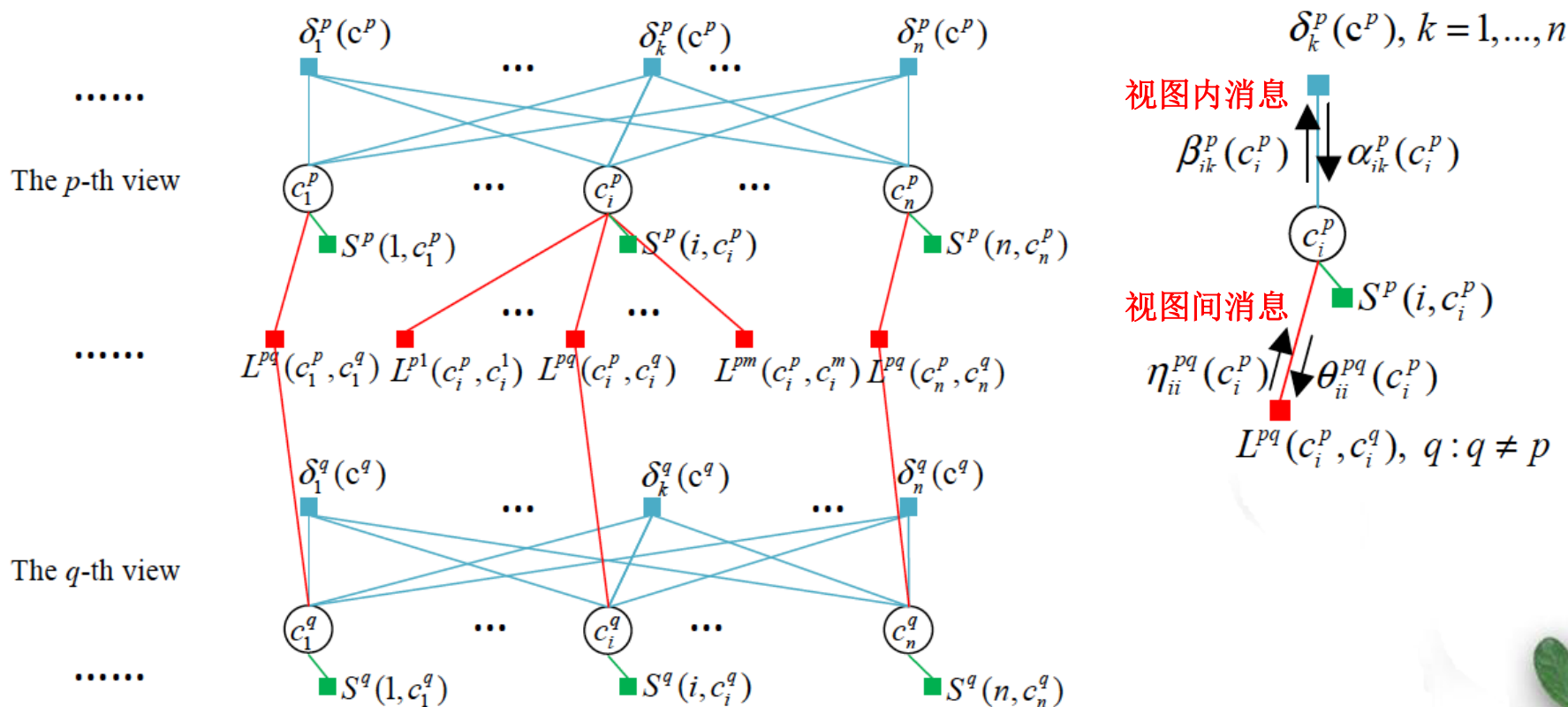


Figure : Factor graph of the global objective function.



多视图相似性传播 (MVAP)

1. Messages passing within the p -th view

$$\tilde{\alpha}_{ik}^p(k) = \begin{cases} \sum_{i': i' \neq k} \max(0, \tilde{\beta}_{i'k}^p(k)) & k = i \\ \min \left[0, \tilde{\beta}_{kk}^p(k) + \sum_{i': i' \notin \{i, k\}} \max(0, \tilde{\beta}_{i'k}^p(k)) \right] & k \neq i \end{cases}$$

$$\tilde{\beta}_{ik}^p(k) = S^p(i, k) + \sum_{q=1, q \neq p}^m \tilde{\eta}_{ii}^{pq}(k) - \max_{j: j \neq k} \left(S^p(i, j) + \tilde{\alpha}_{ij}^p(j) + \sum_{q=1, q \neq p}^m \tilde{\eta}_{ii}^{pq}(j) \right)$$

2. The simplified cross-view messages are as follows: from the q -th view to the p -th view

$$\tilde{\eta}_{ii}^{pq}(k) = \max_{k'} \left[L^{pq}(k, k') + S^q(i, k') + \tilde{\alpha}_{ik'}^q(k') + \sum_{p'=1, p' \neq p, p' \neq q}^m \tilde{\eta}_{ii}^{qp'}(k') \right]$$

The message θ is eliminated.



多视图相似性传播 (MVAP)

- The optimal cluster assignment vectors $\{\hat{\mathbf{c}}^1, \hat{\mathbf{c}}^2, \dots, \hat{\mathbf{c}}^m\}$ are computed as follows,

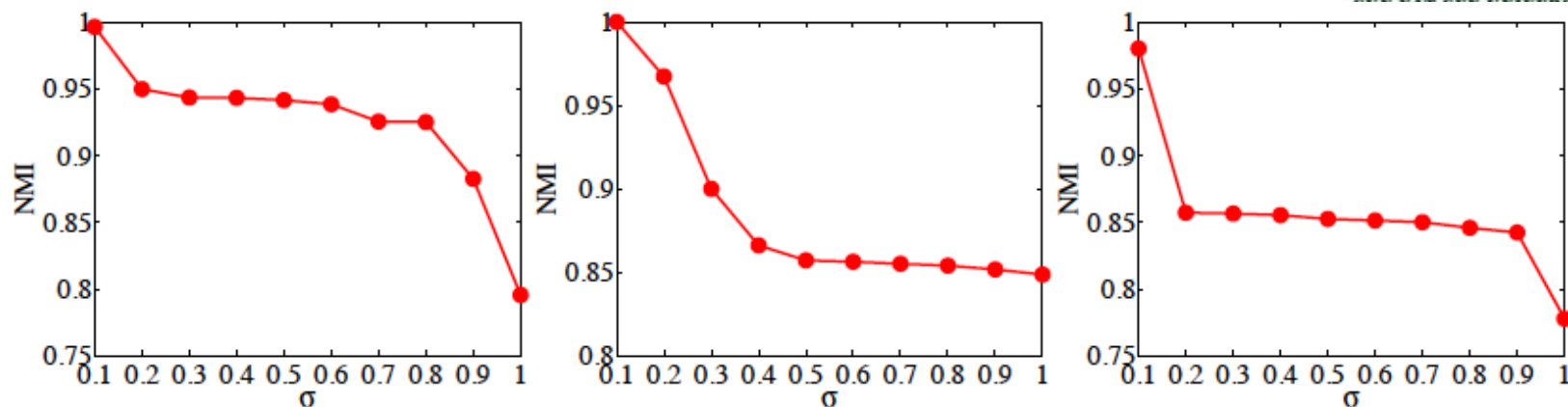
$$\hat{\mathbf{c}}_i^p = \arg \max_j \left[S^p(i, j) + \sum_{q=1, q \neq p}^m \tilde{\eta}_{ij}^{pq}(j) + \tilde{\alpha}_{ij}^p(j) \right]$$

- One common cluster assignment vector $\hat{\mathbf{c}} = [\hat{\mathbf{c}}_1, \hat{\mathbf{c}}_2, \dots, \hat{\mathbf{c}}_n]$ can be obtained where each value $\hat{\mathbf{c}}_i$ maximizes the sum of all incoming messages from all views

$$\hat{\mathbf{c}}_i = \arg \max_j \sum_{p=1}^m \left[S^p(i, j) + \sum_{q=1, q \neq p}^m \tilde{\eta}_{ij}^{pq}(j) + \tilde{\alpha}_{ij}^p(j) \right]$$



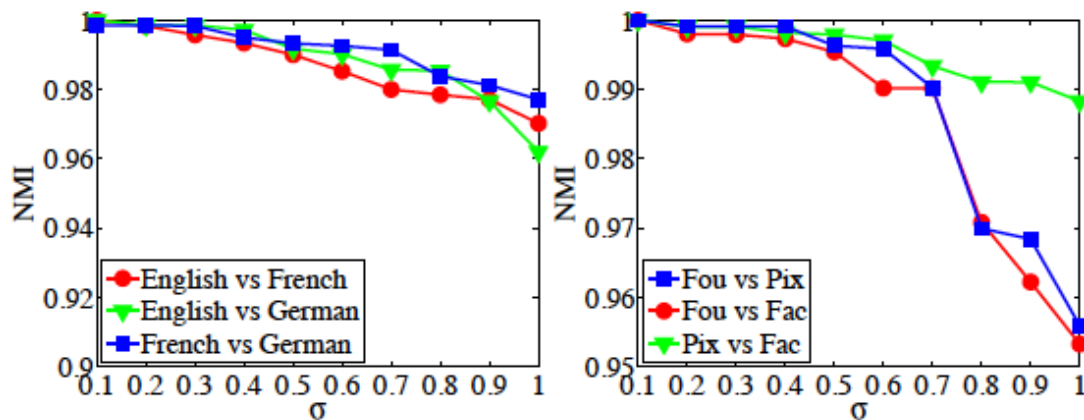
多视图相似性传播 (MVAP)



(a) VIS/NIR dataset

(b) Cora dataset

(c) CiteSeer dataset



(d) Reuters dataset

(e) Mfeat dataset

Figure : Analysis on the trade-off parameter σ w.r.t. the consistency of the clustering results across different views.



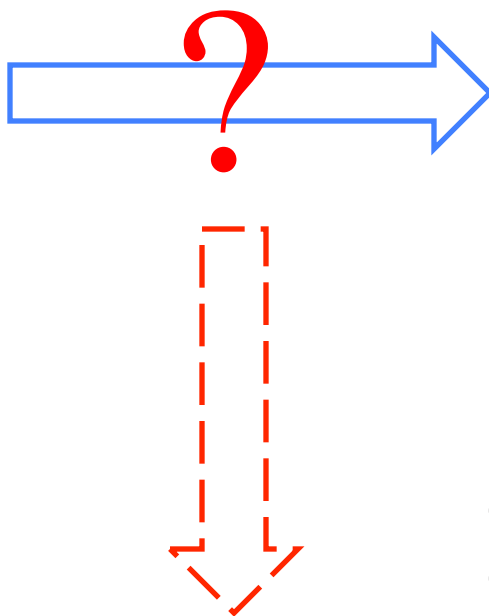
多视图非线性聚类

缺损非线性视图 I

缺损非线性视图 II

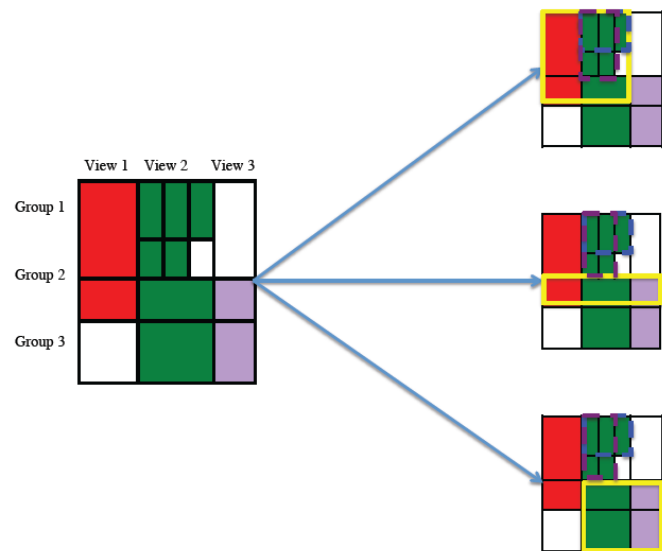
⋮

缺损非线性视图 ...



多视图非线性聚类

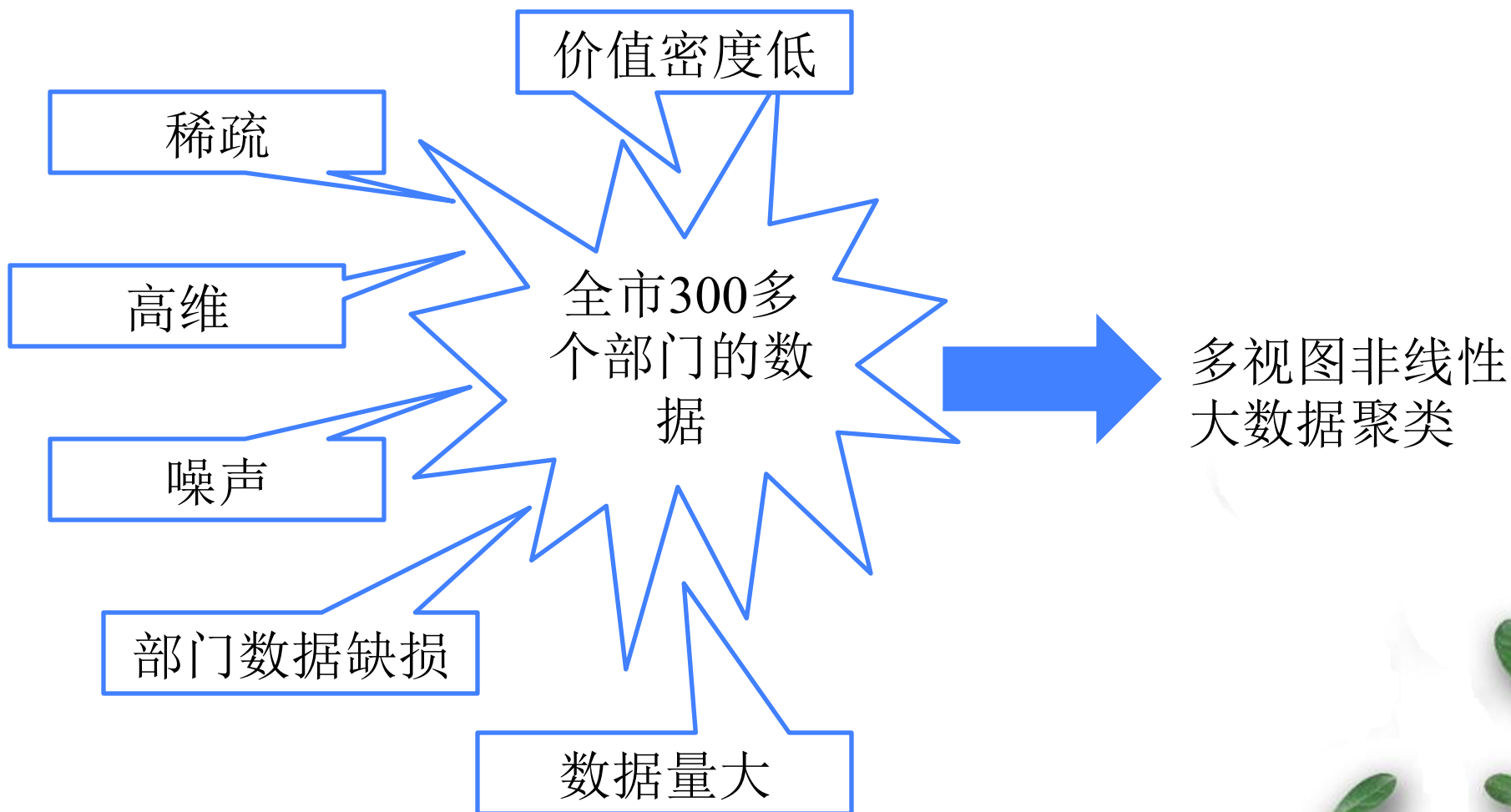
- 1、如何刻画非线性视图？
- 2、非线性视图之间如何进行耦合？
- 3、如何处理视图的缺损空间？





应用案例

● 珠海市公安局的大数据应用之案件聚类分析



Thanks!!!

- ① Chang-Dong Wang, Jian-Huang Lai and Philip S. Yu. Multi-View Clustering Based on Belief Propagation. IEEE Transactions on Knowledge and Data Engineering, Vol. 28, No. 4, pp. 1007-1021, April, 2016.
- ② Chang-Dong Wang, Jian-Huang Lai, Ching Y. Suen and Jun-Yong Zhu. Multi-Exemplar Affinity Propagation. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 35, No. 9, pp.2223-2237, Sept. 2013.