第6章 聚类理论

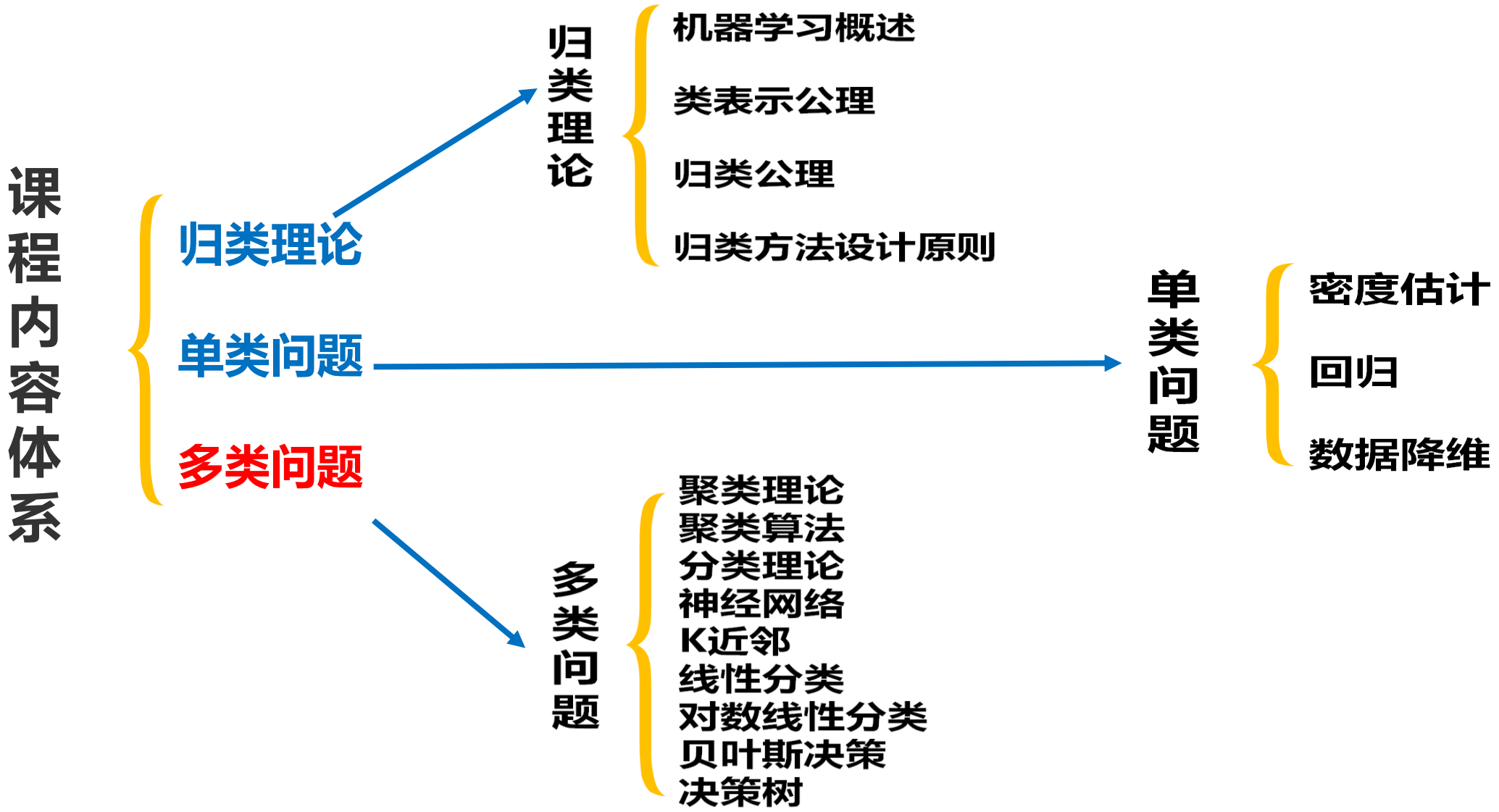
方以类聚，物以群分，吉凶生矣。

——《周易·系辞上》

北京交通大学《机器学习》课程组

目录

课程内容体系

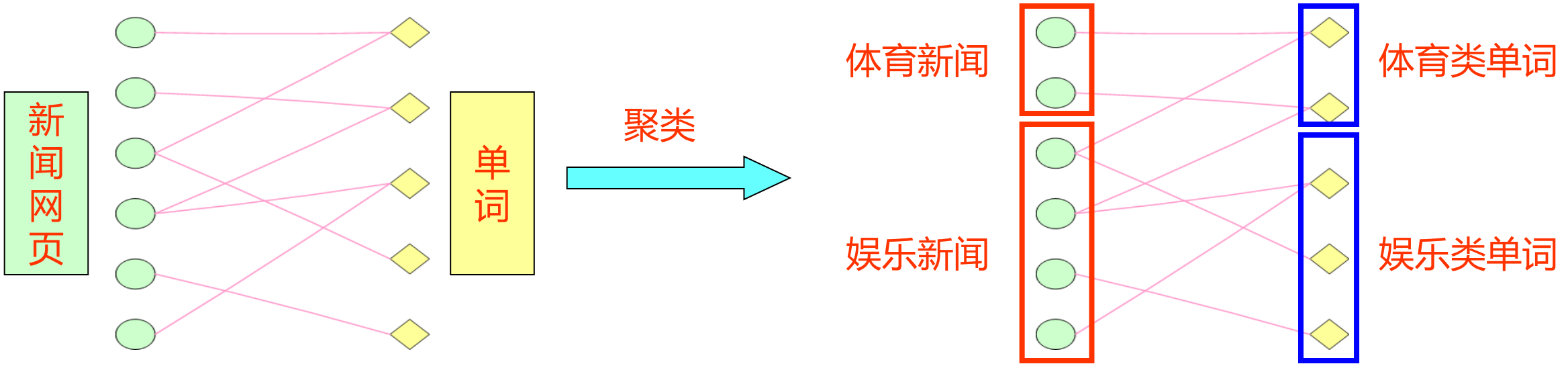

课程内容体系
归类理论
单类问题
多类问题
归类理论
机器学习概述
类表示公理
归类公理
归类方法设计原则
单类问题
密度估计
回归
数据降维
多类问题
聚类理论
聚类算法
分类理论
神经网络
K近邻
线性分类
对数线性分类
贝叶斯决策
决策树

什么是聚类?

- 将有限集合中的对象划分成c个非空子集，使得类内的对象相似、类间的对象不相似
- 训练样本的标记信息未知，无监督学习
- 通过对无标记训练样本的学习来揭示数据的内在性质及规律，为进一步的数据分析提供基础。
- 无监督学习中，应用最广的是“聚类”。

应用举例（1）

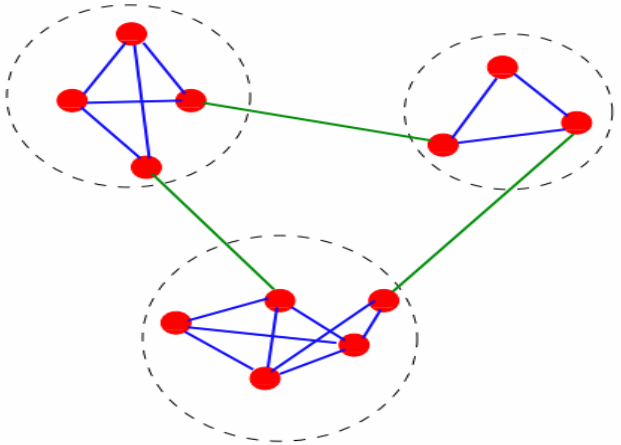
- 文档聚类（topic discovery）

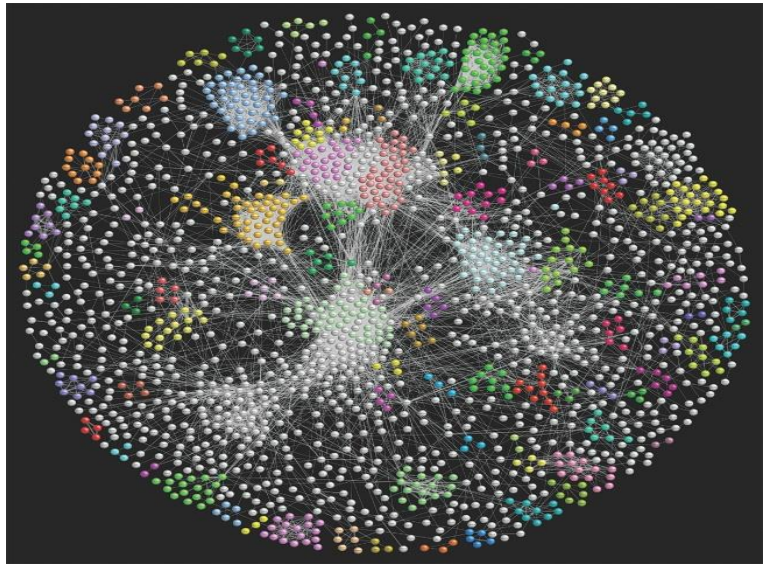


应用举例（2）

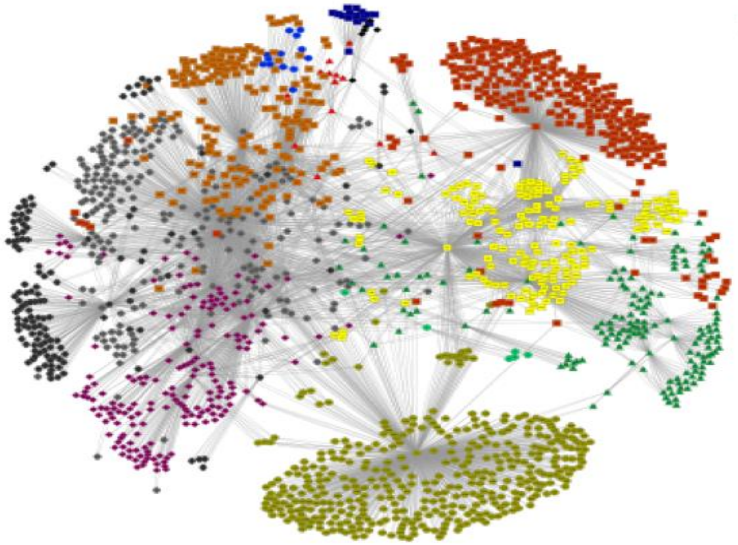
■ 社团发现（community detection）

寻找一个最优划分，使社区内部的连接尽量紧密，社区之间连接尽量稀疏。





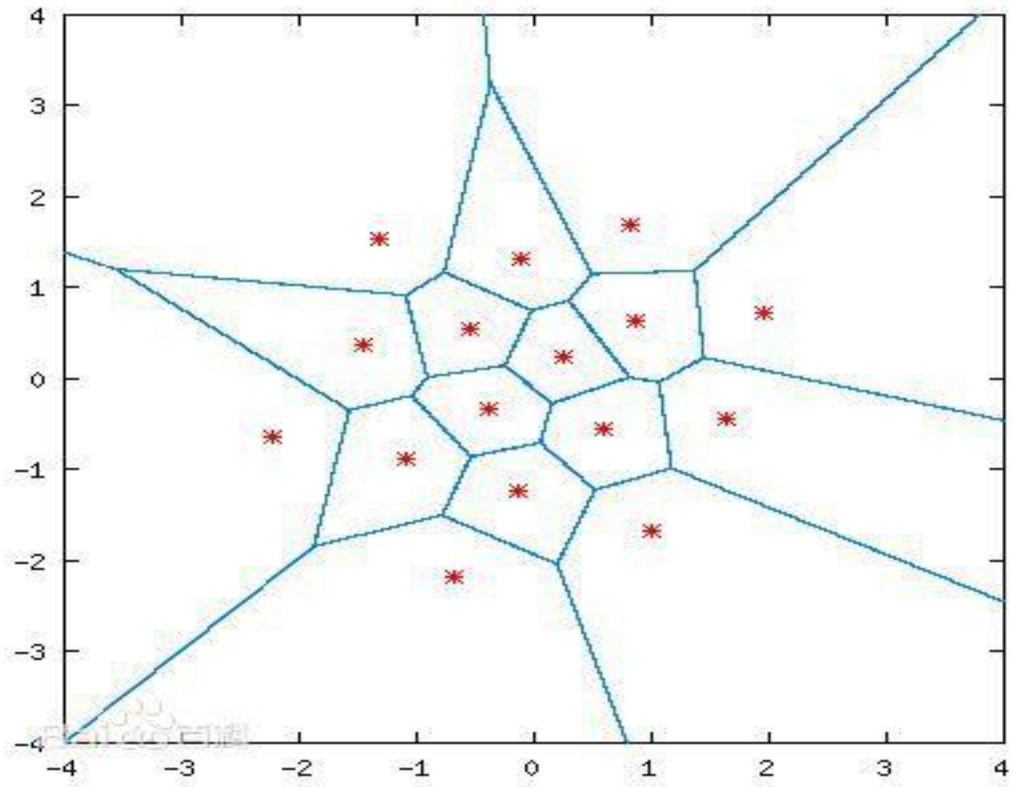
蛋白质作用网络：划分不同的蛋白质功能组



社交网络：划分不同的人群

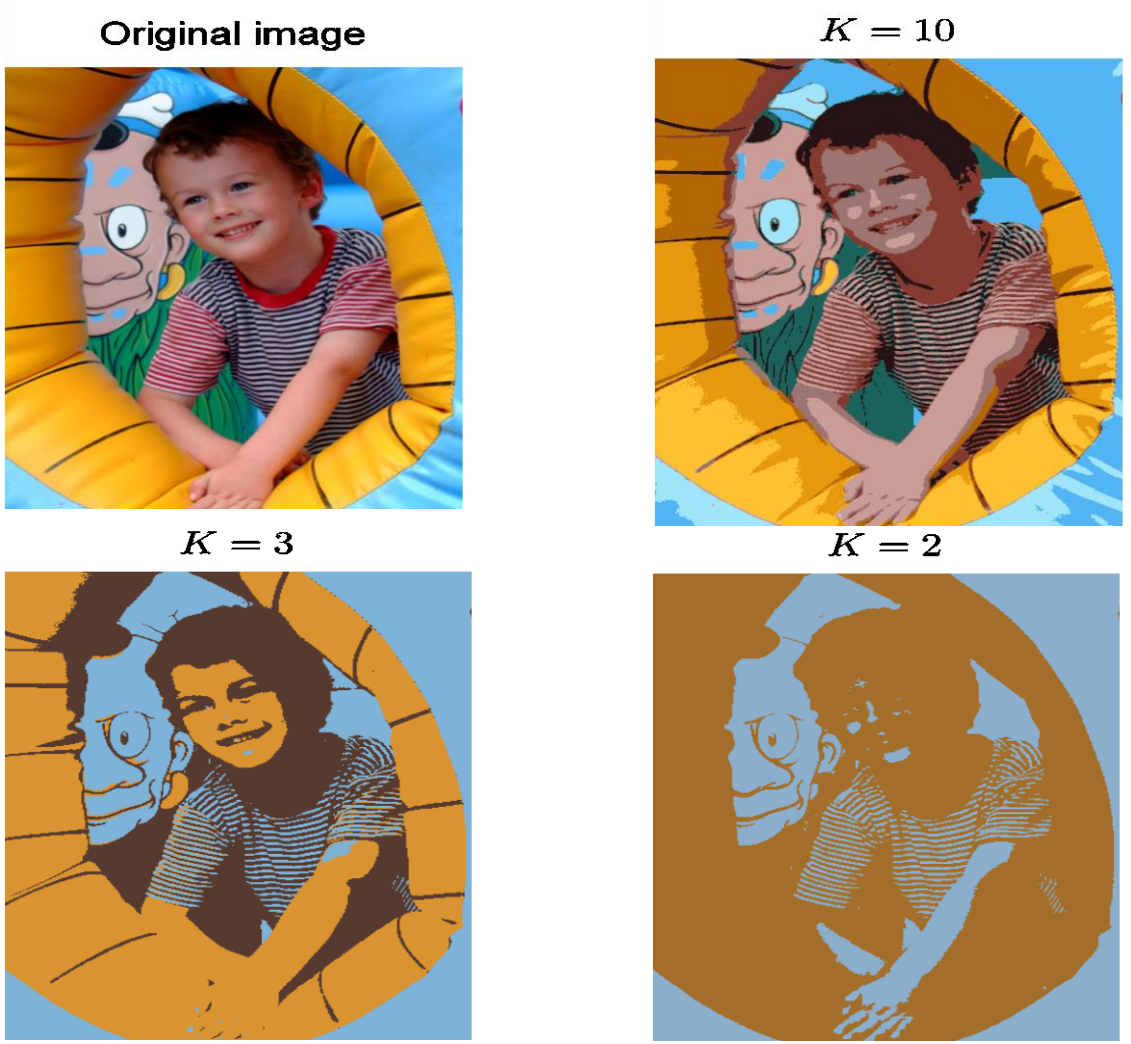
应用举例（3）

- 矢量量化



一种数据压缩技术

将所有像素聚为*K*类



应用举例（4）

- 图像分割

自动驾驶技术中，需使用图像分割技术，将车载摄像头探查到的图像，自动进行分割并归类，以避让行人和车辆等障碍。

聚类的发展

- 思想早
  - “方以类聚，物以群分，吉凶生矣。”《周易. 系辞上》
- 算法晚
  - 上个世纪五十年代， 以层次聚类算法、k-means 算法为代表
- 研究热
  - 聚类分析研究热是本世纪的事情，其标志事件是谱聚类、协同滤波和聚类不可能性定理的提出等

聚类算法的部分理论基础

- Probability theory(Redner & Walker, 1984)
- Information theory(Tishby et al., 1995)
- Graph theory(Wu & Leahy, 1993)
- Fuzzy logic (Bezdek, 1974),
- Matrix decomposition(Xu et al.,2003)
- Game theory(Rota Bulo & Pelillo, 2013)
- Quantum mechanics(Horn & Gottlieb, 2001),
- Gestalt theory(Zahn, 1971)
- Cognitive science(Shepard, Arabie, 1979)

聚类算法的部分代表性文献

- Charu C. Aggarwal, Chandan K. Reddy. DATA CLUSTERING: Algorithms and Applications. CRC Press, Taylor & Francis Group, 2014.

- Brian S. Everitt, Sabine Landau, Morven Leese, Daniel Stahl. Cluster Analysis. John Wiley & Sons, Ltd，2011.

目录

聚类分析的步骤

- 传统的聚类分析方法分为四个部分
  - 数据表示
  - 聚类判据（聚类准则）
  - 聚类算法
  - 聚类评估（聚类有效性评价）

目录

目录

聚类的两个关键问题（1）

- 关键问题1: 何为类?
  - 即类的定义和表示问题

类表示公理

- 7.1 样例理论: 层次聚类算法
  - 凝聚型聚类算法
- 7.2 原型理论: 点原型聚类算法
  - C均值算法
  - 模糊C均值
- 7.3 基于密度估计的聚类算法
  - 基于参数密度估计的聚类算法
    - 基于混合高斯模型的聚类算法
  - 基于无参数密度估计的聚类算法
    - 山峰聚类算法：削峰聚类算法、描峰聚类算法

样本可分性公理与归类理论

▶ **原型理论**

一个对象归为A类而不是其他类，仅仅因为该对象**更像A类的原型表示**而不是其他类的原型表示。

▶ **样例理论**

一个对象归为A类而不是其他类，仅仅因为该对象**更像A类的样例表示**而不是其他类的样例表示。

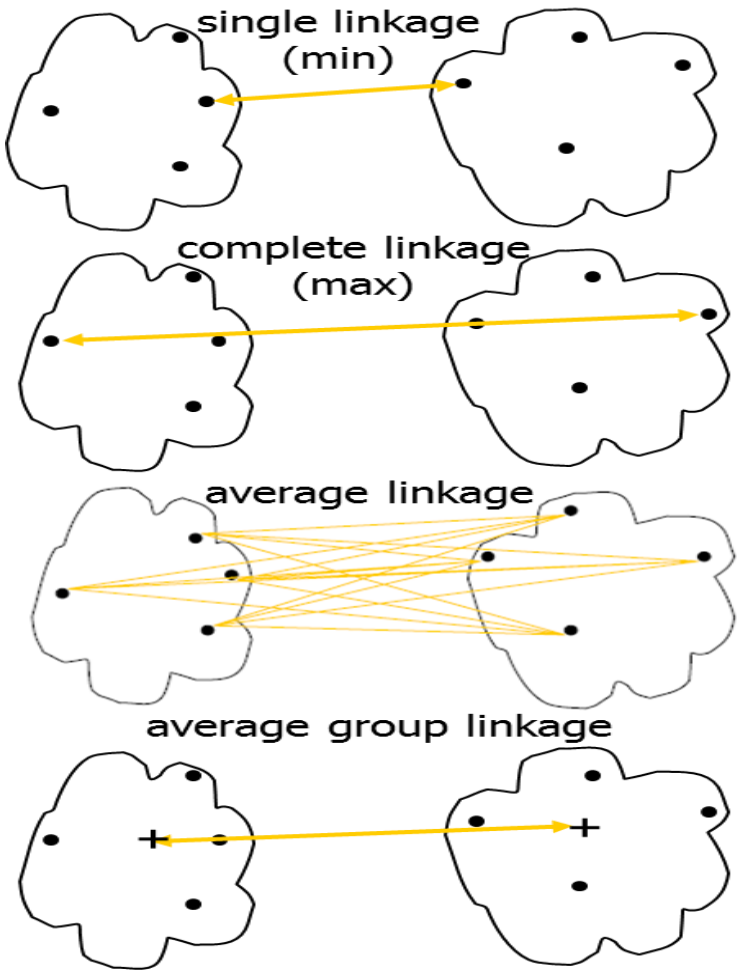
聚类的两个关键问题(2)

归类公理

- 关键问题2：何为类内的对象相似、类间的对象不相似？
  - 即类的相似性计算问题
  - 由样本可分性公理，对象与其对应的类表示最相似，

    →类内的对象应该相似
  - 由类可分性公理，不同的类有不同的类表示

    →类间的对象不应该相似



聚类分析的准则

聚类算法设计准则

- 类紧致性准则: 对象与其所属类具有最大相似度(或最小相异度)
  - 增强样本可分公理的要求
- 类分离性准则: 不同类表示的差异最大(类间距离越大越好)
  - 增强类可分公理的要求
- 类一致性准则: 输入与输出之间误差最小(通常假设为真)
  - 减弱类表示唯一公理的要求
- 奥卡姆剃刀准则: 选择简单的类表示

目录

类紧致准则和聚类不等式

■ **定理6.1** 令$(X, U, \underline{X}, \mathrm{Sim})$表示给定数据集合$X = \{x_1, x_2, \ldots, x_N\}$的聚类结果。如果归类公理成立，则下列不等式成立:

样本与所属类的相似度

样本与任意类的相似度

$$\prod_k \mathrm{Sim}\left(x_k, \underline{X_{\overrightarrow{x_k}}}\right) \geq \prod_k \mathrm{Sim}\left(x_k, \underline{X_{\phi(k)}}\right)$$

$$\sum_k \mathrm{Sim}\left(x_k, \underline{X_{\overrightarrow{x_k}}}\right) \geq \sum_k \mathrm{Sim}\left(x_k, \underline{X_{\phi(k)}}\right)$$

$$\prod_k \mathrm{Sim}\left(x_k, \underline{X_{\overrightarrow{x_k}}}\right) \geq \prod_k \sum_i \alpha_i \mathrm{Sim}\left(x_k, \underline{X_i}\right)$$

$$\sum_k \mathrm{Sim}\left(x_k, \underline{X_{\overrightarrow{x_k}}}\right) \geq \sum_k f\left(\sum_i \alpha_i g\left(\mathrm{Sim}\left(x_k, \underline{X_i}\right)\right)\right)$$

左$=\prod_k \max\left(\mathrm{Sim}\left(x_k, \underline{X_i}\right)\right)$
$=\prod_k (\sum_1^c \alpha_i) \max\left(\mathrm{Sim}\left(x_k, \underline{X_i}\right)\right)$
$= \prod_k \sum_i \alpha_i \max\left(\mathrm{Sim}\left(x_k, \underline{X_i}\right)\right)$
≥右

左$=\sum_k (\sum_i \alpha_i) \mathrm{Sim}\left(x_k, \underline{X_{\overrightarrow{x_k}}}\right)$
$\geq \sum_k (\sum_i \alpha_i) * \max(\mathrm{Sim}(x_k, \underline{X_j})$
$\geq \sum_k \left(\sum_i \alpha_i \mathrm{Sim}\left(x_k, \underline{X_i}\right)\right)$
$= \sum_k \sum_i \alpha_i f\left(g\left(\mathrm{Sim}\left(x_k, \underline{X_i}\right)\right)\right)$
$\geq \sum_k f\left(\sum_i \alpha_i g\left(\mathrm{Sim}\left(x_k, \underline{X_i}\right)\right)\right)$
=右

其中$\phi(k)$是从$\{1,2,\ldots,N\}$到$\{1,2,\ldots,c\}$的函数, $\alpha_i > 0$, $\sum_1^c \alpha_i = 1$; f是凸函数， $\forall t \in R_+, f\left(g(t)\right) = t$。

$$f(\lambda x_1 + (1-\lambda)x_2) \leq \lambda f(x_1) + (1-\lambda) f(x_2)$$

对象与它所属类具有**最大相似度。**

类紧致准则和聚类不等式

- **定理6.2** 令$(X, U, \underline{X}, \mathrm{Ds})$表示给定数据集合$X = \{x_1, x_2, \ldots, x_N\}$的聚类结果。如果归类公理成立，则下列不等式成立:

$$\sum_k \mathrm{Ds}\left(x_k, \underline{X_{\overrightarrow{x_k}}}\right) \leq \sum_k \mathrm{Ds}\left(x_k, \underline{X_{\phi(k)}}\right)$$

$$\prod_k \mathrm{Ds}\left(x_k, \underline{X_{\overrightarrow{x_k}}}\right) \leq \prod_k \mathrm{Ds}\left(x_k, \underline{X_{\phi(k)}}\right)$$

$$\sum_k \mathrm{Ds}\left(x_k, \underline{X_{\overrightarrow{x_k}}}\right) \leq \sum_k f\left(\sum_i \alpha_i g\left(\mathrm{Ds}\left(x_k, \underline{X_i}\right)\right)\right)$$

$$\sum_k \mathrm{Ds}\left(x_k, \underline{X_{\overrightarrow{x_k}}}\right) \leq \prod_k f\left(\sum_i \alpha_i g\left(\mathrm{Ds}\left(x_k, \underline{X_i}\right)\right)\right)$$

其中$\phi(k)$是从$\{1,2,\ldots,N\}$到$\{1,2,\ldots,c\}$的函数, $\alpha_i > 0, \sum_1^c \alpha_i = 1$; f是凹函数，$\forall t \in R_+, f\left(g(t)\right) = t$。

对象与它所属类具有**最小相异度。**

类分离准则和重合类非稳定假设

- **类分离准则：** 加强类可分性要求,不同类表示之间差异越大越好，类间距离越大越好

- 一个目标函数例子

类紧致性准则　　类分离性准则

$$\frac{1}{N}\sum_{i=1}^{c}\sum_{k=1}^{N}u_{ik}^{m}\mathrm{Ds}\left(x_k,\underline{X_i}\right)-\frac{\gamma}{c}\sum_{j=1}^{c}\left\|\underline{X_i}-\underline{X_j}\right\|^2$$

其中$\mathrm{Ds}\left(x_k,\underline{X_i}\right)=\left\|x_k-\underline{X_i}\right\|^2$, $\forall k,\sum_{i=1}^{c}u_{ik}=1, m>1, \gamma>0$。

IEEE TRANSACTIONS ON IMAGE PROCESSING, VOL. 10, NO. 6, JUNE 2001　　923

Fuzzy Algorithms for Combined Quantization and Dithering

Doğan Özdemir, *Member, IEEE,* and Lale Akarun, *Member, IEEE*

类分离准则和重合类非稳定假设

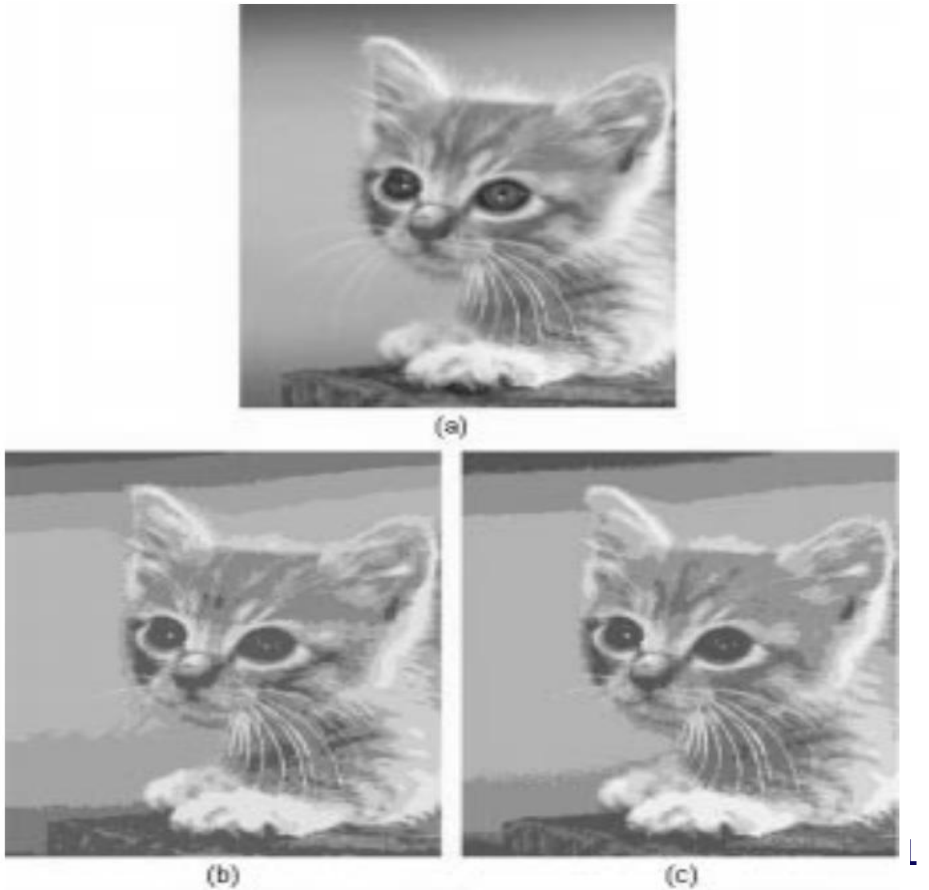
类紧致性准则

类分离性准则

$$\frac{1}{N}\sum_{i=1}^{c}\sum_{k=1}^{N}u_{ik}^{m}\mathrm{Ds}\left(x_k,\underline{X_i}\right)-\frac{\gamma}{c}\sum_{j=1}^{c}\left\|\underline{X_i}-\underline{X_j}\right\|^2$$

原图→8色

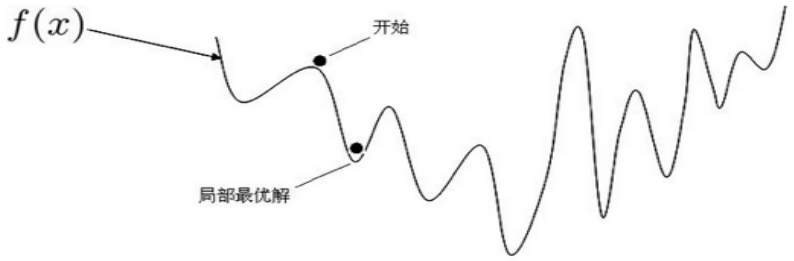
***Abstract*—Color quantization reduces the number of the colors in a color image**, while the subsequent dithering operation attempts to create the illusion of more colors with this reduced palette. In quantization, the palette is designed to minimize the mean squared error (MSE). However, the dithering that follows enhances the color appearance at the expense of increasing the MSE. We introduce three joint quantization and dithering algorithms to overcome this contradiction. The basic idea is the same in two of the approaches: introducing the dithering error to the quantizer in the training phase. **The fuzzy C-means (FCM)** and the fuzzy learning vector quantization (FLVQ) algorithms are used to develop two combined mechanisms. In the third algorithm, we **minimize an objective function including an inter-cluster separation (ICS)** term to obtain a color palette which is more suitable for dithering. The goal is to enlarge the convex hull of the quantization colors to obtain the illusion of more colors after error diffusion. The color contrasts of images are also enhanced with the proposed algorithm. We test the results of these three new algorithms using quality metrics which model the perception of the human visual system and illustrate that substantial improvements are achieved after dithering.

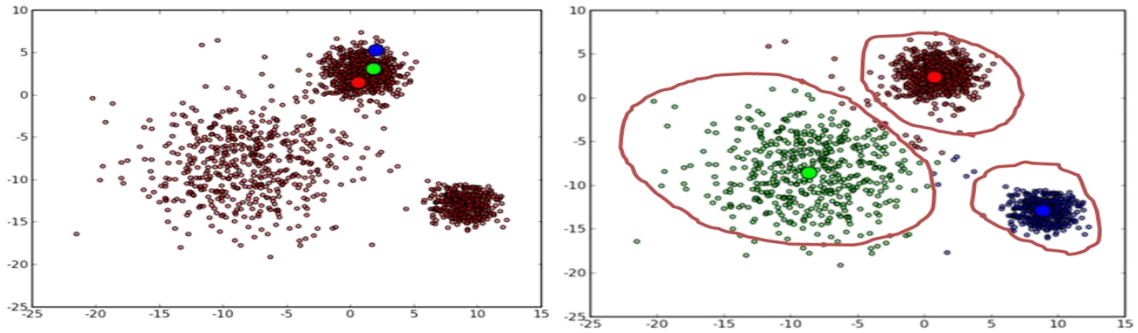


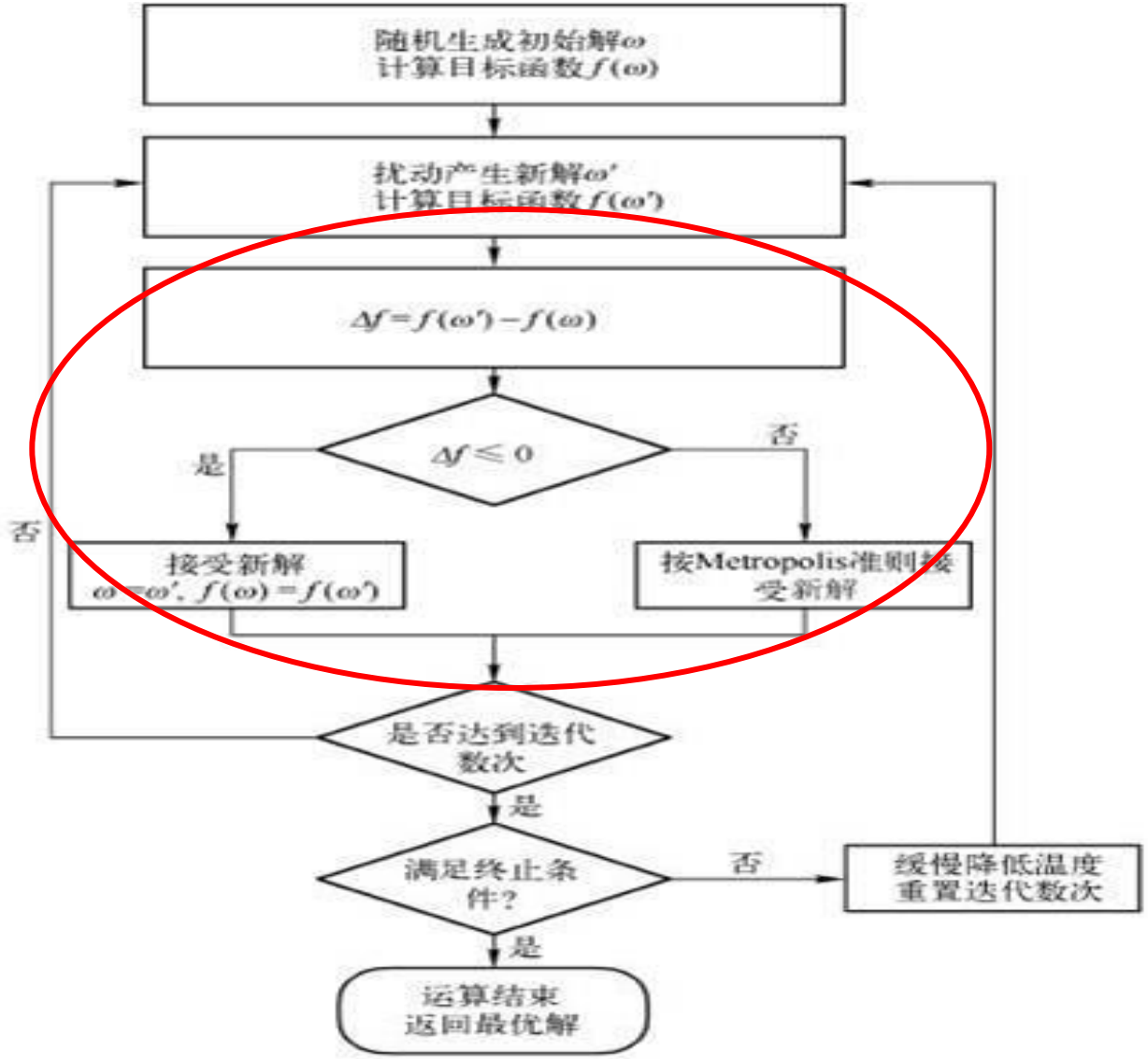
类分离准则和重合类非稳定假设

■重合类非稳定假设

- 如果聚类算法产生了重合聚类结果，则该结果不
- 迭代型聚类算法容易产生重合聚类结果，如决定







类一致性准则和迭代型聚类算法

类一致性准则

- 类表示唯一公理一定成立，考虑类一致性准则
- 类一致性准则：使输入端的外部指称与输出端的外部指称误差最小
- 降低了类表示唯一性公理的要求

迭代型聚类算法:

- 反复迭代输入端的类外延表示和输出端的类内部表示，使其逐步接近类表示唯一性公理

EM算法伪代码

输入: 观察数据$X = \{x_1, x_2, \dots, x_N\}$; 高斯混合模型

输出: $V = [v_{ik}]$; $\underline{Y} = \widehat{\Theta}$

聚类过程:

(1) 初始化参数$\widehat{\Theta}^{(0)} = (\hat{\pi}_1^{(0)}, \hat{\pi}_2^{(0)}, \dots, \hat{\pi}_C^{(0)}, \hat{\theta}_1^{(0)}, \hat{\theta}_2^{(0)}, \dots, \hat{\theta}_C^{(0)})$; 相当于EM算法的E-step

(2) 更新类的外部表示：固定$\widehat{\Theta}$，更新每个样本的隶属度v_{ik}

$$v_{ik}^{(t)} = \frac{\hat{\pi}_i^{(t)} p(x_k|\hat{\theta}_i^{(t)})}{\sum_{i=1}^{C} \hat{\pi}_i^{(t)} p(x_k|\hat{\theta}_i^{(t)})}$$

(3) 更新类的内部表示：固定v_{ik}，更新Θ 相当于EM算法的M-step

$$\hat{\pi}_i^{(t+1)} = \frac{1}{N}\sum_{k=1}^{N} v_{ik}^{(t)} \qquad \hat{\mu}_i^{(t+1)} = \frac{\sum_{k=1}^{N} v_{ik}^{(t)} x_k}{\sum_{k=1}^{N} v_{ik}^{(t)}} \qquad \hat{\Sigma}_i^{(t+1)} = \frac{\sum_{k=1}^{N} v_{ik}^{(t)} \left(x_k - \hat{\mu}_i^{(t)}\right)\left(x_k - \hat{\mu}_i^{(t)}\right)^{\mathrm{T}}}{\sum_{k=1}^{N} v_{ik}^{(t)}}$$

(4) 重复(2)~(3)直到收敛。

聚类算法设计准则

- 类紧致准则和聚类不等式
- 类分离准则和重合类非稳定假设
- 类一致性准则和迭代型聚类算法

目录

聚类的有效性

- 由于U未知，所以聚类是无监督的
- 聚类的目的是发现数据集的隐含结构
- 聚类算法得到的是最佳隐含结构的近似解
- 需要验证聚类结果有效性: 考察聚类结果与真实最佳隐含结构差别有多大

聚类的有效性

- 外部方法
  - 假设数据集已经被标注
  - 比较聚类结果与已知类标的相似程度
  - 一般采用类一致性准则
- 内蕴方法
  - 数据集未标定
  - 采用类分离准则、类紧致准则、奥卡姆剃刀准则
  - 目标是避免重合归类结果和无信息划分

外部方法(1): Rand Index

$$Rand(U,V)=\frac{a_1+a_4}{a_1+a_2+a_3+a_4}$$

√ $a_1=\sum_i\sum_j\frac{n_{ij}(n_{ij}-1)}{2}$,

× $a_2=\sum_j\frac{n_{\cdot j}(n_{\cdot j}-1)}{2}-\sum_i\sum_j\frac{n_{ij}(n_{ij}-1)}{2}$,

× $a_3=\sum_i\frac{n_{i\cdot}(n_{i\cdot}-1)}{2}-\sum_i\sum_j\frac{n_{ij}(n_{ij}-1)}{2}$,

√ $a_4=\frac{N(N-1)}{2}-a1-a2-a3$,

输出

输入	1	2
1	80	20
2	30	70

$a_1=C_{80}^2+C_{20}^2+C_{30}^2+C_{70}^2$

$a_2=C_{110}^2+C_{90}^2-a_1$

$a_3=2*C_{100}^2-a_1$

$a_4=C_{200}^2-a_1-a_2-a_3$

a_1: U和V中均属于同一类的样本对的数目

a_2: U中不属于同一类V中属于同一类的样本对的数目

a_3: U中属于同一类V中不属于同一类的样本对的数目

a_4: U和V中都不属于同一类的样本对的数目

$\mathbf{n_{ij}}$表示在U中分为i类而在V中分为j类的样本数目，$n_{i\cdot}=\sum_j n_{ij}$, $n_{\cdot j}=\sum_i n_{ij}$, $\sum_i\sum_j n_{ij}=N$。

W.M. Rand (1971), "Objective criteria for the evaluation of clustering methods," *Journal of the American Statistical Association*, 66, 846–850.

外部方法(1): Rand Index

尽可能大

$$Rand(U,V)=\frac{a_1+a_4}{a_1+a_2+a_3+a_4}$$

$$\frac{C_N{}^2-a_2-a_3}{C_N{}^2}$$

√ $a_1=\sum_i\sum_j\frac{n_{ij}(n_{ij}-1)}{2}$,

× $a_2=\sum_j\frac{n_{\cdot j}(n_{\cdot j}-1)}{2}-\sum_i\sum_j\frac{n_{ij}(n_{ij}-1)}{2}$,

× $a_3=\sum_i\frac{n_{i\cdot}(n_{i\cdot}-1)}{2}-\sum_i\sum_j\frac{n_{ij}(n_{ij}-1)}{2}$,

√ $a_4=\frac{N(N-1)}{2}-a1-a2-a3$,

$\mathbf{n_{ij}}$表示在U中分为i类而在V中分为j类的样本数目，$n_{i\cdot}=\sum_j n_{ij}$, $n_{\cdot j}=\sum_i n_{ij}$, $\sum_i\sum_j n_{ij}=N$。

W.M. Rand (1971), "Objective criteria for the evaluation of clustering methods," *Journal of the American Statistical Association*, 66, 846–850.

外部方法(2): Normalized Mutual Information

$$H(X) = -\sum_{x \in X} p(x) \log_2 p(x)$$

其中，约定 $0 log\ 0 = 0$。

尽可能大

$$\mathrm{NMI}(U,V) = \frac{2\sum_i \sum_j n_{ij} ln \frac{n_{ij} N}{n_{i.} n_{.j}}}{-\sum_i n_{i.} ln \frac{n_{i.}}{N} - \sum_j n_{.j} ln \frac{n_{.j}}{N}} = 2\left(1 - \frac{H_{ij}}{H_i + H_j}\right)$$

输出

输入	1	2
1	50	50
2	50	50

差的分类

分子：左=0；右=2*2*200*ln(2)

分母左右相等

输出

输入	1	2
1	100	0
2	0	100

好的分类

T. O. Kvalseth (1987), "Entropy and Correlation: Some Comments", IEEE Transactions on Systems Man & Cybernetics, 17(3): 517-519.
http://www.cnblogs.com/gatherstars/p/6004075.html
http://www.cnblogs.com/ziqiao/archive/2011/12/13/2286273.html

内蕴方法

- 设计内蕴聚类有效性指标应遵循的准则
  - 类紧致性、类分离性、极值准则、奥卡姆剃刀准则
- 不同的准则设计不同的聚类有效性指标

内蕴方法

- 划分系数
- 划分熵

$$V_{pc}=\frac{1}{N}\sum_{i=1}^{c}\sum_{k}^{N}u_{ik}^2$$

$$V_{pe}=\frac{1}{N}\sum_{i=1}^{c}\sum_{k}^{N}u_{ik}\ln u_{ik}$$

尽可能大

每个样本的隶属度尽可能晰

基于隶属度定义，符合类紧致性准则

$$u_{ik}=\begin{cases}1, & k=l\\ 0, & k=1\sim c,\ k\neq l\end{cases}$$

$$V_{pc}(\mathrm{i})=\sum_1^c u_{ik}^2=1$$

$$V_{pe}(i)=\sum_{k}^{c}u_{ik}\ln u_{ik}=0$$

$$u_{ik}=\frac{1}{c},\quad k=1\sim c$$

$$V_{pc}(\mathrm{i})=\sum_1^c u_{ik}^2=1/c$$

$$V_{pe}(i)=\sum_{k}^{c}u_{ik}\ln u_{ik}=-\ln(c)$$

内蕴方法(1): Xie-Bein指标

Xie-Bein指标（模糊C均值算法的聚类有效性指标）

$$\mathrm{XB}(X,U,\underline{X})=\frac{\sum_{i=1}^{c}\sum_{k=1}^{N}\left(u_{ik}^{2}\left\|x_k-\underline{X_i}\right\|^{2}\right)}{N\times\min\limits_{i\neq j}\left\|\underline{X_i}-\underline{X_j}\right\|^{2}}$$

类紧致性

类分离性

- 分子表示各个**类的紧致性**，分母表示**类之间的分离度**
- $\mathrm{XB}(X,U,\underline{X})$值越大，聚类结果越差；值越小，聚类结果越好
- **重合划分**会使得$\mathrm{XB}(X,U,\underline{X})$接近无穷大，属于非正则聚类结果

Xie, X. L. and Beni, G. "A validity measure for fuzzy clustering", IEEE Trans. PAMI, 13(8): 841-847,1991.

内蕴方法(2): Davies-Bouldin指标

尽可能小

$$\mathrm{DB}(X, U, \underline{X}) = \frac{1}{Nc}\sum_{i}\max_{j\neq i}\left\{\frac{\sum_{x_k\in X_i} d(x_k, \underline{X_i})}{d(\underline{X_i}, \underline{X_j})\times n_i} + \frac{\sum_{x_k\in X_j} d(x_k, \underline{X_j})}{d(\underline{X_i}, X_j)\times n_j}\right\}$$

类紧致性

类分离性

A Cluster Separation Measure

DAVID L. DAVIES AND DONALD W. BOULDIN

***Abstract*—A measure is presented which indicates the similarity of clusters which are assumed to have a data density which is a decreasing function of distance from a vector characteristic of the cluster.**

The measure can be used to infer the appropriateness of data partitions and can therefore be used to compare relative appropriateness of various divisions of the data. The measure does not depend on either the number of clusters analyzed nor the method of partitioning of the data and can be used to guide a cluster seeking algorithm.

224 IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, VOL. PAMI-1, NO. 2, APRIL 1979

内蕴方法(3): Caliński-Harabasz指标

尽可能大

类分离性

$$\mathrm{CH}(X,U,\underline{X})=\frac{(N-c)\sum_{i=1}^{c}\sum_{k=1}^{N}\left(u_{ik}^{2}\left\|\underline{X_i}-\bar{x}\right\|^{2}\right)}{(c-1)\sum_{i=1}^{c}\sum_{k=1}^{N}\left(u_{ik}^{2}\left\|\underline{X_i}-x_k\right\|^{2}\right)},\bar{x}=\frac{\sum_k x_k}{N}$$

类紧致性

COMMUNICATIONS IN STATISTICS, 3(1), 1-27 (1974)

A DENDRITE METHOD FOR CLUSTER ANALYSIS

T. Caliński and J. Harabasz

Academy of Agriculture, Poznań, Poland

Key Words & Phrases: numerical taxonomy; cluster analysis; minimum variance (WGSS) criterion for optimal grouping; approximate grouping procedure; shortest dendrite = minimum spanning tree; variance ratio criterion for best number of groups.

ABSTRACT

A method for identifying clusters of points in a multi-dimensional Euclidean space is described and its application to taxonomy considered. It reconciles, in a sense, two different approaches to the investigation of the spatial relationships between the points, viz., the agglomerative and the divisive methods. A graph, the shortest dendrite of Florek et al. (1951a), is constructed on a nearest neighbour basis and then divided into clusters by applying the criterion of minimum within-cluster sum of squares. This procedure ensures an effective reduction of the number of possible splits. The method may be applied to a dichotomous division, but is perfectly suitable also for a global division into any number of clusters. An informal indicator of the "best number" of clusters is suggested. It is a "variance ratio criterion" giving some insight into the structure of the points. The method is

BJTU “Machine Learning” Group

于　剑：jianyu@bjtu.edu.cn；　　李晓龙：hlli@bjtu.edu.cn；

景丽萍：lpjing@bjtu.edu.cn；　　吴　丹：wudan@bjtu.edu.cn；

田丽霞：lxtian@bjtu.edu.cn；　　万怀宇：hywan@bjtu.edu.cn；

黄惠芳：hfhuang@bjtu.edu.cn；　　王　晶：wj@bjtu.edu.cn.