# 非监督式机器学习

## 聚类(Clustering)

## --算法评价

河北师范大学软件学院

2018.04.03-04.12

河北师范大学软件学院
Software College of Hebei Normal University

## 评价的意义

(1)避免所发现的数据结构源自噪声干扰

(2)不同聚类算法的比较

(3)两个聚类集合(**two sets of clusters**)的比较

(4)两个聚类的比较

--> 聚类趋向：验证给定数据集是否具有聚类结构；
发现数据中真实的结构

## 评价的几个角度

- 明确给定数据集合中“聚类的趋势”
  如：区分给定数据集内是否存在非随机性“结构”
- “外部评价”--将聚类分析的结果与给定的结果(带有类别标签的专门数据)比较
- “内部评价”--评估聚类分析的结果是否与数据结构相符，而无需参考外部信息- 只借助数据本身
- 比较不同聚类算法的分析结果，以确定哪种聚类算法更好
- 确定正确的“聚类数目”

**评价的几种类型(*types of validation measures*)**

**(1)外部评价(*external validation*)**

需要关于研究对象相关领域的先验知识

如：一个预定义的划分

**不足** 强化了研究者的主观猜测；

会忽略某些与之前认识不符的现象；

导致错过发现新规律新模式的机会

**(2)内部评价(*internal validation*)**

基于数据本身内在的信息，量化分析

# 外部评价的一些常见指标

给定数据集$D=\{x_1,\dots,x_m\}$，$x_i=[x_{i1} \ \cdots \ x_{id}]^T \in R^d$

若 $\begin{cases} \text{由聚类给出的簇划分结果} & C=\{C_1,\dots,C_k\} \\ \text{参考模型给出的簇划分结果} & C^*=\{C_1^*,\dots,C_s^*\} \end{cases}$

数据集$D$内各样本相应簇标记值集合 $\begin{cases} \lambda=\{\lambda_1,\dots,\lambda_m\} \\ \lambda^*=\{\lambda_1^*,\dots,\lambda_m^*\} \end{cases}$

数据集$D$内样本两两配对，定义：

$$a=|SS| \qquad SS=\{(x_i,x_j) \mid \lambda_i=\lambda_j, \lambda_i^*=\lambda_j^*, i<j\}$$

$$b=|SD| \qquad SD=\{(x_i,x_j) \mid \lambda_i=\lambda_j, \lambda_i^*\neq\lambda_j^*, i<j\}$$

$$c=|DS| \qquad DS=\{(x_i,x_j) \mid \lambda_i\neq\lambda_j, \lambda_i^*=\lambda_j^*, i<j\}$$

$$d=|DD| \qquad DD=\{(x_i,x_j) \mid \lambda_i\neq\lambda_j, \lambda_i^*\neq\lambda_j^*, i<j\}$$

显然 $a+b+c+d=C_m^2=\frac{m(m-1)}{2}$

基于上述定义，给出用于聚类性能度量的常见**外部指标**

[1] $Jaccard$系数$(Jaccard\ Coefficient$，简称$JC)$

$$JC=\frac{a}{a+b+c}\in[0,1]$$

[2] $FM$指数$(Fowlkes\ and\ Mallows\ Index$，简称$FMI)$

$$FMI=\sqrt{\frac{a}{a+b}\cdot\frac{a}{a+c}}\in[0,1]$$

[3] $Rand$指数$(Rand\ Index$，简称$RI)$

$$RI=\frac{a+d}{a+b+c+d}=\frac{2(a+d)}{m(m-1)}\in[0,1]$$

# 内部评价的一些常见评价指标

给定数据集$\boldsymbol{D}=\{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_m\}$, $\boldsymbol{x}_i=\begin{bmatrix}x_{i1} & \cdots & x_{id}\end{bmatrix}^T\in\boldsymbol{R}^d$

若由聚类给出的簇划分结果 $\boldsymbol{\mathcal{C}}=\{C_1,\ldots,C_k\}$

$$\text{并且}\begin{cases} dist(\cdot,\cdot)\text{——两样本点之间距离} \\ \mu=\dfrac{1}{|C|}\sum_{x\in C}x\text{——任意簇}C\in\boldsymbol{\mathcal{C}}\text{的中心点.}\end{cases}$$

$\forall C\in\boldsymbol{\mathcal{C}}$, 簇$C$内样本间的平均距离

$$avg(C)=\frac{2}{|C|(|C|-1)}\sum_{1\le i<j\le|C|}dist(x_i,x_j)$$

$\forall C\in\boldsymbol{\mathcal{C}}$, 簇$C$内样本间的最远距离

$$diam(C)=\max_{1\le i<j\le|C|}dist(x_i,x_j)$$

簇$C_i,C_j$样本间最近距离 $d_{\min}(C_i,C_j)=\min_{x_i\in C_i,x_j\in C_j}dist(x_i,x_j)$

簇$C_i,C_j$中心点之间距离 $d_{\text{cen}}(C_i,C_j)=dist(\mu_i,\mu_j)$

基于上述定义，给出用于聚类性能度量的常见**内部指标**

[1] ***DB***系数 (***Davies – Bouldin Index***, 简称***DBI***)

$$DBI = \frac{1}{k}\sum_{i=1}^{k}\max_{j\neq i}\left(\frac{avg(C_i)+avg(C_i)}{d_{\mathrm{cen}}(C_i,C_j)}\right)$$

***DBI***值越小越好.

[2] ***Dunn***指数 (***Dunn Index***, 简称***DI***)

***minimal intercluster distance / maximal intracluster distance.***

$$DI = \min_{1\le i\le k}\min_{j\neq i}\left(\frac{d_{\min}(C_i,C_j)}{\max\limits_{1\le l\le k} diam(C_l)}\right) \qquad DI \in [0,\infty)$$

***DI***值越大越好.