聚类基本问题之性能度量和距离计算

【参考资料】

周志华 《机器学习》

一篇文章透彻解读聚类分析（附数据和R代码）

先给出相关的符号定义。假定样本集$D=\{x_1,x_2,\ldots,x_m\}$包含m个无标记样本，每个样本$\boldsymbol{x}_i=(x_{i1};x_{i2};\ldots;x_{in})$是一个$n$维特征向量，则聚类算法将样本集$D$划分为$k$个不相交的簇$\{C_l|l=1,2;\ldots,k\}$，其中$C_{l'}\cap_{l'\neq l}C_l=\varnothing$且$D=\bigcup_{l=1}^{k}C_l$。相应地，我们用$\lambda_j\in\{1,2,\ldots,k\}$表示样本$\boldsymbol{x}_j$的"簇标记"（cluster label），即$x_j\in C_{\lambda_j}$。于是，聚类的结果可用包含$m$个元素的簇标记向量$\boldsymbol{\lambda}=(\lambda_1;\lambda_2;\ldots;\lambda_m)$表示。

聚类算法涉及到两个基本问题——性能度量和距离计算。

1. 性能度量

聚类的目标是"物以类聚"，即同一簇的样本尽可能彼此相似，不同簇的样本尽可能不同。换言之，聚类结果的"簇内相似度"（intra-cluster similarity）高且"簇间相似度"（inter-cluster similarity）低。

聚类性能度量大致有两类。一类是将聚类结果与某个"参考模型"进行比较，例如领域专家给出的划分结果，这类指标称为"外部指标"（external index）；另一类是直接考察聚类结果而不利用任何参考模型，称为"内部指标"（internal index）。

1.1 外部指标

假设数据集$D=\{x_1,x_2,\ldots,x_m\}$，经过聚类后得到的簇划分为$C=\{C_1,C_2,\ldots,C_s\}$，参考模型给出的簇划分$C^*=\{C_1^*,C_2^*,\ldots,C_s^*\}$，相应的，令$\lambda$和$\lambda^*$分别表示与$C$和$C^*$对应的簇标记向量。我们将样本两两配对考虑，定义：

$$\begin{aligned}
a&=|SS|,SS=\left\{(x_i,x_j)\,|\lambda_i=\lambda_j,\lambda_i^*=\lambda_j^*,i<j\right\}\\
b&=|SD|,SD=\left\{(x_i,x_j)\,|\lambda_i=\lambda_j,\lambda_i^*\neq\lambda_j^*,i<j\right\}\\
c&=|DS|,DS=\left\{(x_i,x_j)\,|\lambda_i\neq\lambda_j,\lambda_i^*=\lambda_j^*,i<j\right\}\\
d&=|DD|,DD=\left\{(x_i,x_j)\,|\lambda_i\neq\lambda_j,\lambda_i^*\neq\lambda_j^*,i<j\right\}
\end{aligned}\tag{1.1}$$

其中集合SS包含了在C中隶属于相同簇且在C^*中也隶属于相同簇的样本对，集合SD包含了在C中隶属于相同簇但在C^*中隶属于不同簇的样本对，其他以此类推。由于每个样本对$(\boldsymbol{x}_i,\boldsymbol{x}_j)\ (i<j)$仅能出现在一个集合中，因此有$a+b+c+d=m(m-1)/2$成立。

基于式（1.1）可以导出聚类性能度量常用的外部指标：

- Jaccard系数（Jaccard Coefficient，JC）

$$\mathrm{JC}=\frac{a}{a+b+c}$$

- FM指数（Fowlkes and Mallows Index，FMI）

$$\mathrm{FMI}=\sqrt{\frac{a}{a+b}\cdot\frac{a}{a+c}}$$

- Rand指数（Rand Index，RI）

$$\mathrm{RI}=\frac{2(a+d)}{m(m-1)}$$

显然，上述性能度量的结果均在$[0,1]$区间内，且值越大越好。

1.2 内部指标

考虑聚类结果的簇划分$\mathcal{C}=\{C_1,C_2,\ldots,C_k\}$，定义

$$\operatorname{avg}(C)=\frac{2}{|C|(|C|-1)}\sum_{1\leqslant i<j\leqslant|C|}\operatorname{dist}(\boldsymbol{x}_i,\boldsymbol{x}_j) \tag{1.2}$$

$$\operatorname{diam}(C)=\max_{1\leqslant i<j\leqslant|C|}\operatorname{dist}(\boldsymbol{x}_i,\boldsymbol{x}_j) \tag{1.3}$$

$$d_{\min}\left(C_i,C_j\right)=\min_{\boldsymbol{x}_i\in C_i,\boldsymbol{x}_j\in C_j}\operatorname{dist}(\boldsymbol{x}_i,\boldsymbol{x}_j) \tag{1.4}$$

$$d_{\text{cen}}\left(C_i,C_j\right)=\operatorname{dist}(\boldsymbol{\mu}_i,\boldsymbol{\mu}_j) \tag{1.5}$$

其中，$\operatorname{dist}(\cdot,\cdot)$用于计算样本之间的距离；$\mu$代表簇$C$的中心点$\mu=\frac{1}{|C|}\sum_{1\leq i\leq|C|}x_i$。显然，$\operatorname{avg}(C)$对应于簇$C$内样本间的平均距离，$\operatorname{diam}(C)$对应于簇$C$内样本间的最远距离，$d_{\min}\left(C_i,C_j\right)$对应于簇$C_i$与簇$C_j$最近样本间的距离，$d_{\text{cen}}\left(C_i,C_j\right)$对应于簇$C_i$与簇$C_j$中心点间的距离。

基于式（1.2）~（1.5）可以导出常用的聚类性能度量指标：

- DB指数（Davies-Bouldin Index，DBI）

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

- Dunn指数（Dunn Index，DI）

$$\mathrm{DI}=\min_{1\leq i\leq k}\left\{\min_{j\neq i}\left(\frac{d_{\min}\left(C_i,C_j\right)}{\max_{1\leq l\leq k}\operatorname{diam}(C_l)}\right)\right\}$$

显然，DBI的值越小越好，而DI则相反，值越大越好。

2. 距离计算

根据属性的数值类型不同，需要使用不同的距离计算方式，下面列举以下常见类型属性的距离计算方法。

2.1 数值属性

数值属性（numerical attribute），又称连续属性（continuous attribute），即取值连续的属性。给定样本$\boldsymbol{x}_i=(x_{i1};x_{i2};\ldots;x_{in})$与样本$\boldsymbol{x}_j=(x_{j1};x_{j2};\ldots;x_{jn})$，最常用的是**闵科夫斯基距离**（Minkowski Distance）：

$$\mathrm{dist}_{\mathrm{mk}}(\boldsymbol{x}_i,\boldsymbol{x}_j)=\left(\sum_{u=1}^{n}|x_{iu}-x_{ju}|^p\right)^{\frac{1}{p}}$$

$p=2$时，闵科夫斯基距离即**欧式距离**（Euclidean Distance）：

$$\mathrm{dist}_{\mathrm{ed}}(\boldsymbol{x}_i,\boldsymbol{x}_j)=\|\boldsymbol{x}_i-\boldsymbol{x}_j\|_2=\sqrt{\sum_{u=1}^{n}|x_{iu}-x_{ju}|^2}$$

$p=1$时，闵科夫斯基距离即曼哈顿距离（Manhattan Distance）：

$$\mathrm{dist}_{\mathrm{man}}(\boldsymbol{x}_i,\boldsymbol{x}_j)=\|\boldsymbol{x}_i-\boldsymbol{x}_j\|_1=\sum_{u=1}^{n}|x_{iu}-x_{ju}|$$

当样本空间中不同属性的重要性不同时，可使用“加权距离”，如带权闵科夫斯基距离：

$$\mathrm{dist}_{\mathrm{wmk}}(\boldsymbol{x}_i,\boldsymbol{x}_j)=(w_1\cdot|x_{i1}-x_{j1}|^p+\ldots+w_n\cdot|x_{in}-x_{jn}|^p)^{\frac{1}{p}}$$

需要注意的是，在计算连续属性之间的距离时，需要考虑数据标度的问题，比如某个属性取值范围是（2000，3000），另一个属性的取值范围是（10，20）。这时需要对数据进行标准化，比如Z-score标准化：

$$Z_f=\frac{X_f-mean_f}{S_f}$$

2.2 二值属性

即取值为0或1的属性，可以利用**列联表**（contingency table）和**Jaccard相似度**计算距离。我们通过一个例子来说明二值属性的距离计算。

假设我们有三个同学，他们有不同的特征，我们想衡量他们哪一对特征是更接近的：

Name	Gender	Fever	Cough	Test-1	Test-2	Test-3	Test-4
Jack	M	Yes	No	Pos	Neg	Neg	Neg
Mary	F	Yes	No	Pos	Neg	Pos	Neg
Jim	M	Yes	Yes	Neg	Neg	Neg	Pos

我们首先将变量用0，1表示

Name	Fever	Cough	Test-1	Test-2	Test-3	Test-4
Jack	1	0	1	0	0	0
Mary	1	0	1	0	1	0
Jim	1	1	0	0	0	1

对于样本X和Y，建立如下的列联表

		Y		
		1	0	sum
X	1	a	b	a+b
	0	c	d	c+d
	sum	a+c	b+d	p

其中a代表的是X和Y中取值都为1的属性数，其他以此类推。我们使用Jaccard相似度来计算距离，如下：

$$d(X,Y)=\frac{b+c}{a+b+c}$$

Jaccard相似度实际上就是交并比，这里改变了它的原始定义（原来分母上应为a），因为需要遵循“距离越大相似度越小”的原则。$b+c$代表的是样本X和样本Y中，其中一方取1，另一方取0的情况。

于是，我们可以计算这三位同学的距离：

$$d(Jack,Mary)=\frac{0+1}{2+0+1}=0.33$$

$$d(Jack,Jim)=\frac{1+2}{1+1+2}=0.75$$

$$d(Mary,Jim)=\frac{2+2}{1+2+2}=0.8$$

所以Jack和Mary是最相近的两个。

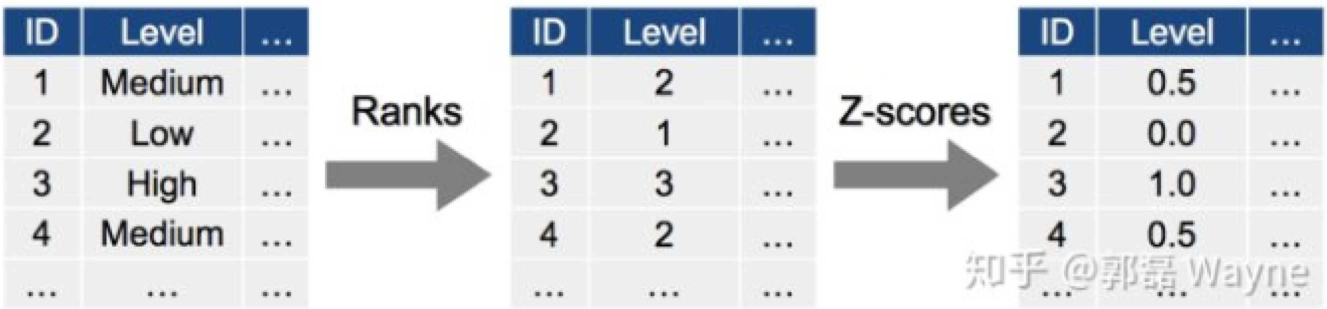
2.3 有序属性

有序属性（ordinal attribute）的取值通常是离散的，但是不同的取值间有顺序关系，比如Level∈{Low, Medium, High}。计算距离的方法是转化为连续变量，再用闵科夫斯基距离计算，即：

1）用每个值对应的排名$r\in[1\ldots N]$来代替这个值；

2）计算z-scores来标准化排名，让r在$[0,1]$之间；

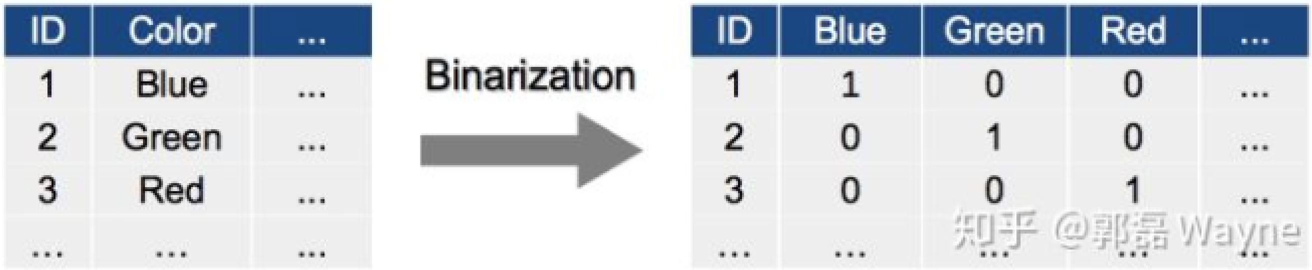
3）计算闵科夫斯基距离。



2.4 无序属性

无序属性，或者说类别属性，取值是离散的，例如Color∈{blue, green, red, ...}。这里介绍两种处理方法。

第一种是将类别属性二值化，比如



ID	Color	...
1	Blue	...
2	Green	...
3	Red	...
...	...	...

ID	Blue	Green	Red	...
1	1	0	0	...
2	0	1	0	...
3	0	0	1	...
...	...	...	...	...

转化成二值属性之后再列联表分析。

第二种是采用**VDM**（Value Difference Metric）。令$m_{u,a}$表示在属性u上取值为a的样本数，$m_{u,a,i}$表示在第i个样本簇中在属性u上取值为a的样本数，k为样本簇数，则属性u上两个离散值a与b之间的VDM距离为

$$\mathrm{VDM}_p(a,b)=\sum_{i=1}^{k}\left|\frac{m_{u,a,i}}{m_{u,a}}-\frac{m_{u,b,i}}{m_{u,b}}\right|^p$$

使用VDM的另一个好处是，可以结合闵科夫斯基距离处理混合属性。假定有n_c个有序属性、$n-n_c$个无序属性，不失一般性，令有序属性排列在无序属性之前，则

$$\mathrm{MinkovDM}_p(\boldsymbol{x}_i,\boldsymbol{x}_j)=\left(\sum_{u=1}^{n_c}|x_{iu}-x_{ju}|^p+\sum_{u=n_c+1}^{n}\mathrm{VDM}_p(x_{iu},x_{ju})\right)^{\frac{1}{p}}$$