第九章 聚类

9.1 聚类

9.1.1 无监督学习

在**"无监督学习"** (unsupervised learning) 中, 训练样本的**标记信息是未知**的, **目标**是通过**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. 此类学习任务中研究最多、应用最广的是**"聚类"** (clustering).

9.1.2 聚类中的簇

聚类试图将数据集中的样本划分为**若干个**通常是**不相交的子集**, **每个子集**称为**一个"簇"** (cluster). 通过这样的划分, 每个簇可能对应于一些潜在的概念(如类别). 同时, 也要注意, 这些概念对聚类算法而言事先是未知的, **聚类过程仅能自动形成簇结构**, 簇所对应的概念语义需由**使用者来把握和命名**.

9.1.3 簇的数学表示

假定样本集 $D=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\ldots,\boldsymbol{x}_m\}$ 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 $\boldsymbol{x}_j\in C_{\lambda_j}$. 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.

聚类既能作为一个**单独过程**, 用于找寻数据内在的分布结构, 也可作为分类等**其他学习任务的前驱过程**.

9.2 性能度量

9.2.1 聚类性能度量相关基本概念

1 聚类性能度量

- 聚类性能度量亦称聚类"**有效性指标**" (validity index). 与监督学习中的性能度量作用相似, 对聚类结果, 我们需通过某种**性能度量来评估其好坏**.
- 另一方面, 若明确了最终将要使用的性能度量, 则可直接**将其作为聚类过程的优化目标**, 从而更好地得副符合要求的聚类结果.

2 簇的"物以类聚"

聚类是将样本集 D 划分为若干互不相交的子集, 即样本簇.

好的聚类结果是"**物以类聚**". 也即是**同一簇的样本尽可能彼此相似, 不同簇的样本尽可能不同**.

- **"簇内相似度"高**
- **"簇间相似度"低**

3 聚类性能度量的分类

一类是将聚类结果与某个"**参考模型**" (reference model) 进行比较, 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).

9.2.2 聚类性能度量的数学表示

1 聚类性能度量的外部指标

对数据集 $D=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\ldots,\boldsymbol{x}_m\}$, 假定通过聚类给出的簇划分为 $\mathcal{C}=\{C_1,C_2,\ldots,C_k\}$, 参考模型给出的簇划分为 $\mathcal{C}^*=\{C_1^*,C_2^*,\ldots,C_s^*\}$. 相应的, 令 $\boldsymbol{\lambda}$ 与 $\boldsymbol{\lambda}^*$ 分别表示与 C 和 C^* 对应的簇标记向量. 将样本两两配对, 定义如下:

$$a=|SS|,\quad SS=\left\{(\boldsymbol{x}_i,\boldsymbol{x}_j)\,|\lambda_i=\lambda_j,\lambda_i^*=\lambda_j^*,i<j\Big)\right\} \tag{9.1}$$

$$b=|SD|,\quad SD=\left\{(\boldsymbol{x}_i,\boldsymbol{x}_j)\,|\lambda_i=\lambda_j,\lambda_i^*\neq\lambda_j^*,i<j\Big)\right\} \tag{9.2}$$

$$c=|DS|,\quad DS=\left\{(\boldsymbol{x}_i,\boldsymbol{x}_j)\,|\lambda_i\neq\lambda_j,\lambda_i^*=\lambda_j^*,i<j\Big)\right\} \tag{9.3}$$

$$d=|DD|,\quad DD=\left\{(\boldsymbol{x}_i,\boldsymbol{x}_j)\,|\lambda_i\neq\lambda_j,\lambda_i^*\neq\lambda_j^*,i<j\Big)\right\} \tag{9.4}$$

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, ...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.

> **注1**: **关于** $a+b+c+d=m(m-1)/2$ **的推导**
>
> 当 $i=1$ 时, j 可以取 $m-1$ 个; 当 $i=2$ 时, j 可以取 $m-2$ 个; ....;当 $i=m-1$ 时, j 可以取 1 个.
>
> 所以也就是 $m-1,m-2,\ldots\ldots,2,1$ 等差数列. 求和有 $[(m-1)+1]*(m-1)/2$, 也即是 $m(m-1)/2$.

基于式 $(9.1)\sim(9.4)$ 可导出下面这些常用的聚类性能度量**外部指标**:

- Jaccard 系数 (简称 JC)

$$\mathrm{JC}=\frac{a}{a+b+c} \tag{9.5}$$

- FM 指数 (简称 FMI)

$$\mathrm{FMI}=\sqrt{\frac{a}{a+b}\cdot\frac{a}{a+c}} \tag{9.6}$$

- Rand 指数 (简称 RI)

$$\mathrm{RI}=\frac{2(a+d)}{m(m-1)} \tag{9.7}$$

上述的三个性能度量的结果**值均在 $[0,1]$ 区间, 且值越大越好.**

2 聚类性能度量的内部指标

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

$$\operatorname{avg}(C)=\frac{2}{|C|(|C|-1)} \sum_{1 \leqslant i<j \leqslant|C|} \operatorname{dist}\left(\boldsymbol{x}_i, \boldsymbol{x}_j\right) \tag{9.8}$$

$$\operatorname{diam}(C)=\max _{1 \leqslant i<j \leqslant|C|} \operatorname{dist}\left(\boldsymbol{x}_i, \boldsymbol{x}_j\right) \tag{9.9}$$

$$d_{\min }\left(C_i, C_j\right)=\min _{\boldsymbol{x}_i \in C_i, \boldsymbol{x}_j \in C_j} \operatorname{dist}\left(\boldsymbol{x}_i, \boldsymbol{x}_j\right) \tag{9.10}$$

$$d_{\text {cen }}\left(C_i, C_j\right)=\operatorname{dist}\left(\boldsymbol{\mu}_i, \boldsymbol{\mu}_j\right) \tag{9.11}$$

其中, $\operatorname{dist}(\cdot,\cdot)$ 用于计算两个样本之间的距离; $\boldsymbol{\mu}$ 代表簇 C 的中心点 $\boldsymbol{\mu}=\frac{1}{|C|} \sum_{1 \leqslant i \leqslant|C|} \boldsymbol{x}_i$.

则 $\operatorname{avg}(C)$ 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, $d_{cen}(C_i,C_j)$ 对应于簇 C_i 与簇 C_j 中心点的距离.

基于式 $(9.8) \sim (9.11)$ 可导出下面这些常用的聚类性能度量内部指标:

- DB 指数 (简称 DBI)

$$\mathrm{DBI}=\frac{1}{k} \sum_{i=1}^k \max _{j \neq i}\left(\frac{\operatorname{avg}\left(C_i\right)+\operatorname{avg}\left(C_j\right)}{d_{\text {cen }}\left(C_i, C_j\right)}\right) \tag{9.12}$$

- Dunn 指数 (简称 DI)

$$\mathrm{DI}=\min _{1 \leqslant i \leqslant k}\left\{\min _{j \neq i}\left(\frac{d_{\min }\left(C_i, C_j\right)}{\max _{1 \leqslant l \leqslant k} \operatorname{diam}\left(C_l\right)}\right)\right\} \tag{9.13}$$

显然, DBI 的值越小越好, 而 DI 则相反, 值越大越好.

> **注2: 关于式 (9.12) 的理解**
>
> 该式的 DBI 值越小越好, 因为我们希望“物以类聚”, 即同一簇的样本尽可能彼此相似, $\operatorname{avg}(C_i)$ 和 $\operatorname{avg}(C_j)$ 越小越好; 我们希望不同簇的样本尽可能不同, 即 $d_{\text{cen}}\left(C_i, C_j\right)$ 越大越好。
>
> 同时, 书上印刷有误, 应该将分母 $d_{\text{cen}}\left(\boldsymbol{\mu}_i, \boldsymbol{\mu}_j\right)$ 改为 $d_{\text{cen}}\left(C_i, C_j\right)$.

9.3 距离计算

9.3.1 距离度量的基本性质

对函数 $\operatorname{dist}(\cdot,\cdot)$, 若它是一个**"距离对量"**(distance measure), 则需满足一些**基本性质**:

- 非负性:

$$\operatorname{dist}\left(\boldsymbol{x}_i, \boldsymbol{x}_j\right) \geqslant 0 \tag{9.14}$$

- 同一性:

$$\operatorname{dist}\left(\boldsymbol{x}_i, \boldsymbol{x}_j\right)=0 \text{当且仅当} \boldsymbol{x}_i=\boldsymbol{x}_j \tag{9.15}$$

- 对称性:

$$\operatorname{dist}(\boldsymbol{x}_i, \boldsymbol{x}_j) = \operatorname{dist}(\boldsymbol{x}_j, \boldsymbol{x}_i) \tag{9.16}$$

- 直递性:

$$\operatorname{dist}(\boldsymbol{x}_i, \boldsymbol{x}_j) \leqslant \operatorname{dist}(\boldsymbol{x}_i, \boldsymbol{x}_k) + \operatorname{dist}(\boldsymbol{x}_k, \boldsymbol{x}_j) \tag{9.17}$$

9.3.2 闵科夫斯基距离

给定样本 $\boldsymbol{x}_i = (x_{i1}; x_{i2}; \ldots; x_{in})$ 与 $\boldsymbol{x}_j = (x_{j1}; x_{j2}; \ldots; x_{jn})$, 最常用的是 "闵科夫斯基距离):

$$\operatorname{dist}_{\mathrm{mk}}(\boldsymbol{x}_i, \boldsymbol{x}_j) = \left(\sum_{u=1}^{n} |x_{iu} - x_{ju}|^p\right)^{\frac{1}{p}} \tag{9.18}$$

对 $p \geqslant 1$, 式 (9.18) 满足距离度量的四个基本性质的.

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

$$\operatorname{dist}_{\mathrm{ed}}(\boldsymbol{x}_i, \boldsymbol{x}_j) = \|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2 = \sqrt{\sum_{u=1}^{n} |x_{iu} - x_{ju}|^2} \tag{9.19}$$

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

$$\operatorname{dist}_{\mathrm{man}}(\boldsymbol{x}_i, \boldsymbol{x}_j) = \|\boldsymbol{x}_i - \boldsymbol{x}_j\|_1 = \sum_{u=1}^{n} |x_{iu} - x_{ju}| \tag{9.20}$$

9.3.3 连续属性和离散属性

1 连续属性和离散属性的定义

连续属性" (continuous attribute)是在定义域上有无穷多个可能的取值，而"离散属性" (categorical attribute) 是在定义域上是有限个取值.

2 距离计算中"序"的概念

在讨论距离计算时, 属性上是否定义了**"序"**关系更为重要.

- **有序属性:** 具有"序"的离散属性与连续属性的性质更接近一些, 能直接在属性值上计算距离. 这种属性称为有序属性.
- **无序属性:** 而不具有"序"这种概念的离散属性, 则不能直接在属性值上计算距离, 这种属性称为无序属性.

闵科夫斯基距离可用与有序属性

3 无序属性的距离计算(VDM)

无序属性距离的计算:

对**无序属性**可采用 **VDM** (Value Difference Metric) 计算距离.

令 $m_{u,a}$ 表示属性 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:

$$\mathrm{VDM}_p(a, b) = \sum_{i=1}^{k} \left|\frac{m_{u,a,i}}{m_{u,a}} - \frac{m_{u,b,i}}{m_{u,b}}\right|^p \tag{9.21}$$

3 无序和有序混合属性的距离计算:

将闵可夫斯基距离和 VDM 结合即可处理混合属性. 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

$$\operatorname{MinkovDM}_p(\boldsymbol{x}_i,\boldsymbol{x}_j)=\left(\sum_{u=1}^{n_c}|x_{iu}-x_{ju}|^p+\sum_{u=n_c+1}^{n}\operatorname{VDM}_p(x_{iu},x_{ju})\right)^{\frac{1}{p}} \tag{9.22}$$

4 属性重要性不同时距离的计算:

当样本空间中**不同属性的重要性不同**时, 可使用**"加权距离"**.

以加权闵科夫斯基距离为例:

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

其中权重 $w_i \geqslant 0(i=1,2,\ldots,n)$ 表示不同属性的重要性, 且通常 $\sum_{i=1}^n w_i=1$

9.4 原型聚类

9.4.1 k 均值算法

给定样本集 $D=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\ldots,\boldsymbol{x}_m\}$, "k 均值"($k-means$)算法针对据类所得簇划分 $\mathcal{C}=\{C_1,C_2,\ldots,C_k\}$ 最小化平方误差

$$E=\sum_{i=1}^{k}\sum_{\boldsymbol{x}\in C_i}\|\boldsymbol{x}-\boldsymbol{\mu}_i\|_2^2 \tag{9.24}$$

其中 $\boldsymbol{\mu}_i=\frac{1}{|C_i|}\sum_{\boldsymbol{x}\in C_i}\boldsymbol{x}$ 是簇 C_i 的均值向量.

式 (9.24) 在一定程度上刻画了**簇内样本围绕簇均值向量的紧密程度** , E **值越小则簇内样本相似度越高.**

输入: 样本集 $D=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\ldots,\boldsymbol{x}_m\}$;
聚类簇数 k.

过程:

从 D 中随机选择 k 个样本作为初始均值向量 $\{\boldsymbol{\mu}_1,\boldsymbol{\mu}_2,\ldots,\boldsymbol{\mu}_k\}$
repeat
  令 $C_i=\varnothing\ (1\leqslant i\leqslant k)$
  for $j=1,2,\ldots,m$ **do**
    计算样本 $\boldsymbol{x}_j$ 与各均值向量 $\boldsymbol{\mu}_i\ (1\leqslant i\leqslant k)$ 的距离: $d_{ji}=||\boldsymbol{x}_j-\boldsymbol{\mu}_i||_2$;
    根据距离最近的均值向量确定 $\boldsymbol{x}_j$ 的簇标记: $\lambda_j=\arg\min_{i\in\{1,2,\ldots,k\}} d_{ji}$;
    将样本 $\boldsymbol{x}_j$ 划入相应的簇: $C_{\lambda_j}=C_{\lambda_j}\bigcup\{\boldsymbol{x}_j\}$;
  end for
  for $i=1,2,\ldots,k$ **do**
    计算新均值向量: $\boldsymbol{\mu}_i'=\frac{1}{|C_i|}\sum_{\boldsymbol{x}\in C_i}\boldsymbol{x}$;
    if $\boldsymbol{\mu}_i'\neq\boldsymbol{\mu}_i$ **then**
      将当前均值向量 $\boldsymbol{\mu}_i$ 更新为 $\boldsymbol{\mu}_i'$
    else
      保持当前均值向量不变
    end if
  end for
until 当前均值向量均未更新

输出: 簇划分 $\mathcal{C}=\{C_1,C_2,\ldots,C_k\}$

图 9.2 k 均值算法