курс «Машинное обучение»

Качество кластеризации

Александр Дьяконов

27 февраля 2023 года

Оценка результатов кластеризации

Если знаем верную кластеризацию… внешняя оценка (External evaluation)

Вопрос: когда?

ничего не знаем $\Rightarrow$ согласованность с данными

внутренняя оценка (Internal evaluation)

Оценка результатов кластеризации: «Internal evaluation»

Пусть чёткая (нет пересечений) кластеризация $U = u_1 \cup \ldots \cup u_{|U|}$

множества $X = \{x_1, \ldots, x_m\}$

Davies–Bouldin index

Использует центроиды и дисперсии

$$DB = \frac{1}{|U|} \sum_{i=1}^{|U|} \max_{j \neq i} \left(\frac{\sigma_i + \sigma_j}{d(c_i, c_j)} \right)$$

Dunn index =

min между кластерами / max внутри кластерами

$$D = \frac{\min\limits_{1 \le i < j \le |U|} d(u_i, u_j)}{\max d_{\text{in}}(u_i)}$$

Silhouette

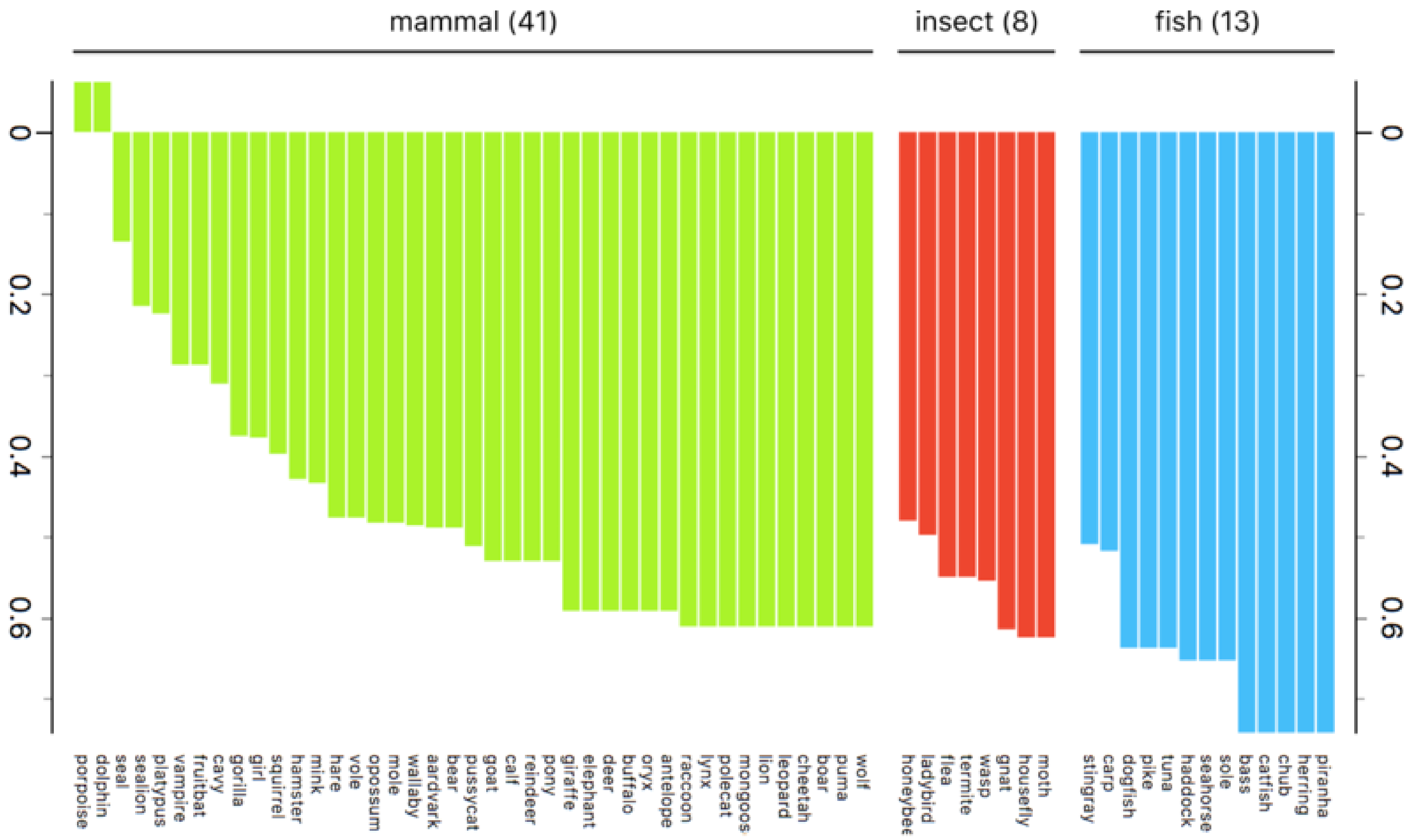
$$x_i \in u_1,\ d(x_i, u_2) \le d(x_i, u_3) \le \ldots$$

Расстояние считается как среднее до всех точек кластера

$$\text{silhouette}(x_i) = \frac{d(x_i, u_2) - d(x_i, u_1)}{\max(d(x_i, u_2), d(x_i, u_1))}$$

Можно усреднять по точкам

Оценка результатов кластеризации: «Internal evaluation»



https://en.wikipedia.org/wiki/Silhouette_(clustering)

Calinski-Harabasz Index (Variance Ratio Criterion)

$$\frac{\operatorname{trace}\left(\frac{1}{|U|-1}\sum_{i=1}^{|U|}|U_i|(x-c_i)(x-c_i)^{\mathrm{T}}\right)}{\operatorname{trace}\left(\frac{1}{m-|U|}\sum_{i=1}^{|U|}\sum_{x\in U_i}(x-c_i)(x-c_i)^{\mathrm{T}}\right)}$$

след матрицы межклассовой ковариации / след матрицы внутриклассовой ковариации

лучше подходит для выпуклых кластеров и евклидовой метрики

External evaluation: взаимная информация

Пусть чёткие (нет пересечений) кластеризации

$$U = u_1 \cup \ldots \cup u_{|U|}$$

$$V = v_1 \cup \ldots \cup v_{|V|}$$

множества $X = \{x_1, \ldots, x_m\}$

$$p_i = \frac{|u_i|}{m}$$

$$H(U) = -\sum_{i=1}^{|U|} p_i \log p_i$$

Аналогично $H(V)$

$$p_{ij} = \frac{|u_i \cap v_j|}{m}$$

$$MI = \sum_{i=1}^{|U|} \sum_{j=1}^{|V|} p_{ij} \log \frac{p_{ij}}{p_i p_j}$$

потом MI ~ насколько более чётко определена U при знании V
уже её можно использовать…

External evaluation: нормализованная взаимная информация

Normalized Mutual Information

$$NMI(U,V) = \frac{MI(U,V)}{\mathrm{mean}(H(U),H(V))}$$

External evaluation: скорректированная взаимная информация

Adjusted mutual information

$$AMI(U,V)=\frac{MI(U,V)-\mathbf{E}(MI(U,V))}{\max(H(U),H(V))-\mathbf{E}(MI(U,V))}$$

1 – если кластеризации равны

~0 – если кластеризации случайны

матожидание можно вычислить аналитически

нужно калибровать, т.к. чем больше кластеров в кластеризациях, тем больше значение MI

```
from sklearn.metrics import mutual_info_score # MI
from sklearn.metrics import normalized_mutual_info_score # [0, 1]
from sklearn.metrics.cluster import adjusted_mutual_info_score
adjusted_mutual_info_score([0, 0, 1, 1], [0, 0, 1, 1])
```

https://en.wikipedia.org/wiki/Adjusted_mutual_information

External evaluation: V-мера

V – среднее гармоническое homogeneity и completeness

homogeneity ~ каждый кластер содержит только объекты отдельного класса

completeness ~ все объекты конкретного класса отнесены в один кластер

$$h = 1 - \frac{H(V\,|\,U)}{H(V)}$$

$$c = 1 - \frac{H(U\,|\,V)}{H(U)}$$

```
from sklearn.metrics.cluster import homogeneity_score

from sklearn.metrics.cluster import completeness_score

from sklearn.metrics.cluster import v_measure_score

v_measure_score([0, 0, 1, 1], [0, 0, 1, 1])
```

External evaluation: Adjusted Rand index

Аналогичная «Adjusted» идея, но проще…

поскольку кластеризация задаёт отношение эквивалентности

Rand index

$$R=\frac{|\{i,j:(i\sim_U j)\,\&\,(i\sim_V j)\}|+|\{i,j:(i\not\sim_U j)\,\&\,(i\not\sim_V j)\}|}{C_m^2}$$

теперь калибровка под случайную кластеризацию:

$X \diagdown Y$	Y_1	Y_2	$\ldots$	Y_s	Sums
X_1	n_{11}	n_{12}	$\ldots$	n_{1s}	a_1
X_2	n_{21}	n_{22}	$\ldots$	n_{2s}	a_2
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
X_r	n_{r1}	n_{r2}	$\ldots$	n_{rs}	a_r
Sums	b_1	b_2	$\ldots$	b_s	

$$\overbrace{ARI}^{\text{Adjusted Index}}=\frac{\overbrace{\sum_{ij}\binom{n_{ij}}{2}}^{\text{Index}}-\overbrace{\left[\sum_i\binom{a_i}{2}\sum_j\binom{b_j}{2}\right]/\binom{n}{2}}^{\text{Expected Index}}}{\underbrace{\frac{1}{2}\left[\sum_i\binom{a_i}{2}+\sum_j\binom{b_j}{2}\right]}_{\text{Max Index}}-\underbrace{\left[\sum_i\binom{a_i}{2}\sum_j\binom{b_j}{2}\right]/\binom{n}{2}}_{\text{Expected Index}}}$$

```
from sklearn.metrics.cluster import adjusted_rand_score
adjusted_rand_score([0, 0, 1, 1], [0, 0, 1, 1])
```

https://en.wikipedia.org/wiki/Rand_index#Adjusted_Rand_index

External evaluation: общий подход

Кластеризация ~ классификация пар

$$\{x_1,\ldots,x_m\}\to\{(1,1),\ldots,(i,j),\ldots,(m,m)\}$$

$$a_U(i,j)=1\Leftrightarrow i\sim_U j$$

Можно сравнивать классификации a_U **и** a_V

Пример, Rand index: $\mathrm{RI}=\dfrac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{FP}+\mathrm{FN}+\mathrm{TN}}$

Fowlkes-Mallows index (FMI)

– среднее геометрическое точности и полноты

$$\mathrm{FMI}=\frac{\mathrm{TP}}{\sqrt{(\mathrm{TP}+\mathrm{FP})(\mathrm{TP}+\mathrm{FN})}}$$

```
from sklearn.metrics.cluster import fowlkes_mallows_score
fowlkes_mallows_score([0, 0, 1, 1], [0, 0, 1, 1])
```

есть и много других…

Литература

К.Д. Маннинг, П. Рагхаван, Х. Шютце «Введение в информационный поиск». — Вильямс, 2011.