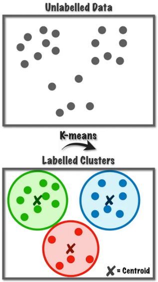
Оценка качества кластеризации

Виктор Китов

victorkitov.github.io



Курс поддержан фондом 'Интеллект'

Победитель конкурса VK среди курсов по IT

Оценка качества кластеризации

Оценка качества кластеризации:

- если кластеризация-промежуточный этап: по качеству итоговой задачи
- если нет разметки:
  - используют идею, что кластеризация хороша, если:
    - объекты одного кластера похожи
    - объекты разных кластеров непохожи
- если есть разметка:
  - учитывать инвариантность к переименованию
  - имеет смысл для малого #размеченных объектов
    - иначе - классификация

Содержание

Метрики качества[1]

- Пусть z_n - номер кластера для x_n.
- Среднее внутрикластерное расстояние:

$$F_0 = \frac{\sum_{i<j} \mathbb{I}\left[z_i = z_j\right] \rho\left(x_i, x_j\right)}{\sum_{i<j} \mathbb{I}[z_i = z_j]}$$

- Среднее межкластерное расстояние:

$$F_1 = \frac{\sum_{i<j} \mathbb{I}\left[z_i \neq z_j\right] \rho\left(x_i, x_j\right)}{\sum_{i<j} \mathbb{I}[z_i \neq z_j]}$$

- Композитные метрики:

$$F_0/F_1,\ F_1 - F_0$$

[1] Какие метрики нужно максимизировать, а какие - минимизировать?

Индекс Дэвиса-Болдуина

- $s_i = \frac{1}{|C_i|}\sum_{n\in C_i}\rho\left(\mu_i, x_n\right)$ - радиус кластера i.
- $d_{ij} = \rho\left(\mu_i, \mu_j\right)$ - расстояние между центроидами i и j.
- Качество разделения кластеров i и j:

$$R_{ij} = \frac{s_i + s_j}{d_{ij}}$$

- Индекс Дэвиса-Болдуина:

$$DB = \frac{1}{K}\sum_{k=1}^{K}\max_{i\neq k} R_{ik}$$

$\oplus$: Быстро вычисляется.

$\ominus$: Поощряет выпуклые кластера

$\ominus$: тип расстояния определяет μ_i

Коэффициент силуэта[2]

Качество кластеризации каждого объекта x_i:

$$Silhouette_i = \frac{d_i - s_i}{\max\{d_i, s_i\}}$$

где среднее расстояние от x_i до объектов

- s_i - того же кластера
- d_i - ближайшего чужого кластера

Общее качество классификации (коэффициент силуэта):

$$Silhouette = \frac{1}{N}\sum_{i=1}^{N}\frac{d_i - s_i}{\max\{d_i, s_i\}}$$

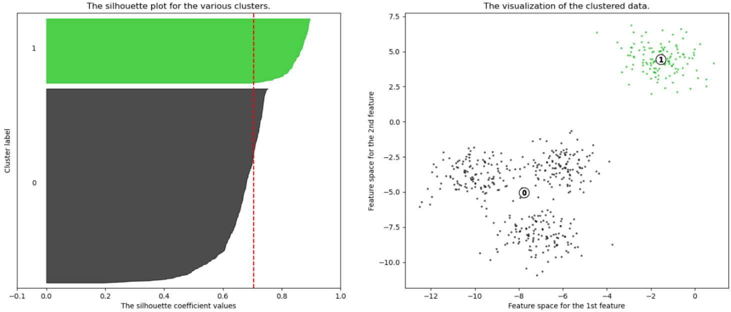
[2] Rousseeuw (1987). "Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis". Computational and Applied Mathematics 20: 53–65.

Обсуждение

- Преимущества
  - Интерпретируемость: $Silhouette \in [-1, 1]$,
    - 1: идеальная кластеризация
    - 0: случайная кластеризация
    - -1: полностью некорректная (инвертированная) кластеризация
- Недостатки
  - сложность $O(N^2 D)$
    - можно рассчитывать по случайной подвыборке
  - поощряет выпуклые кластеры

Подбор #кластеров по силуэту[3]

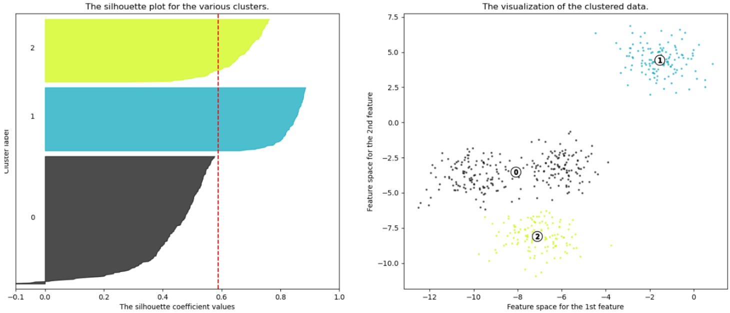
- Отсортируем объекты в каждом кластере по коэффициенту силуэта.
- Качество кластеризации - среднее значение коэффициента и отсутствие отрицательных значений.



[3] Эксперимент в sklearn.

Подбор #кластеров по силуэту[3]

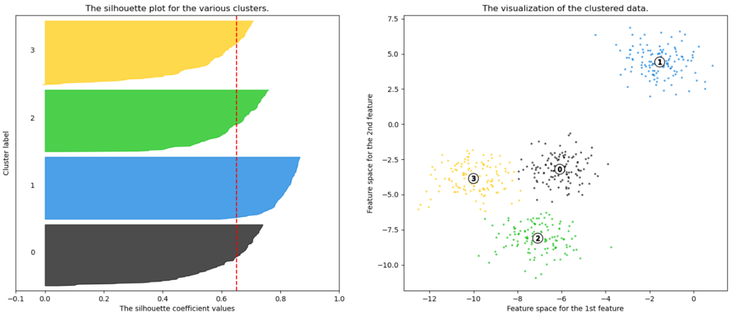
- Отсортируем объекты в каждом кластере по коэффициенту силуэта.
- Качество кластеризации - среднее значение коэффициента и отсутствие отрицательных значений.



[3] Эксперимент в sklearn.

Подбор #кластеров по силуэту[3]

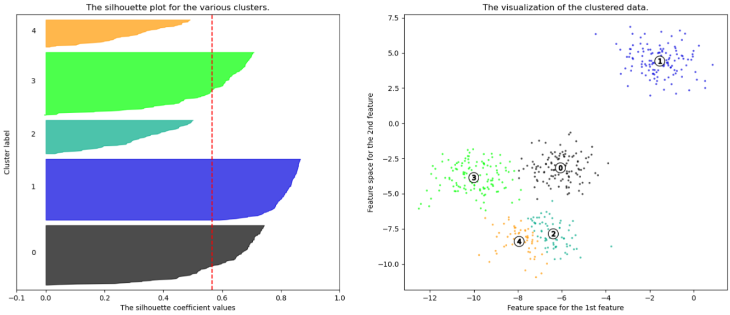
- Отсортируем объекты в каждом кластере по коэффициенту силуэта.
- Качество кластеризации - среднее значение коэффициента и отсутствие отрицательных значений.



[3] Эксперимент в sklearn.

Подбор #кластеров по силуэту[3]

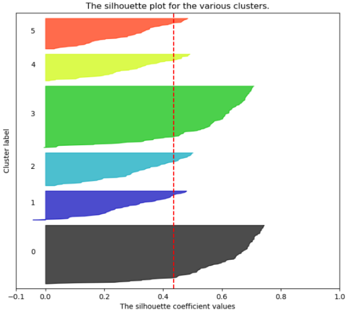
- Отсортируем объекты в каждом кластере по коэффициенту силуэта.
- Качество кластеризации - среднее значение коэффициента и отсутствие отрицательных значений.

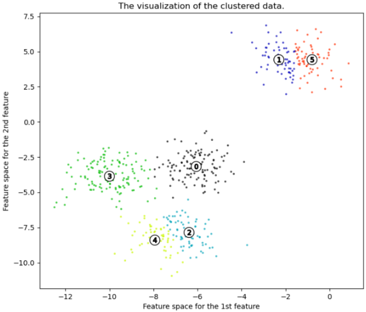


[3] Эксперимент в sklearn.

Подбор #кластеров по силуэту[3]

- Отсортируем объекты в каждом кластере по коэффициенту силуэта.
- Качество кластеризации - среднее значение коэффициента и отсутствие отрицательных значений.





[3] Эксперимент в sklearn.

Индекс Калинского[4]

- Внутрикластерная (within cluster) ковариационная матрица

$$W = \frac{1}{N-K}\sum_{k=1}^{K}\sum_{x\in C_k}(x-\mu_k)(x-\mu_k)^T$$

- Межкластерная (between cluster) ковариационная матрица

$$B = \frac{1}{K-1}\sum_{k=1}^{K}N_k(\mu_k-\mu)(\mu_k-\mu)^T$$

- Индекс Калинского:

$$I = \frac{\mathrm{tr}\, B}{\mathrm{tr}\, W} = \frac{N-K}{K-1}\frac{\mathrm{tr}\left\{\sum_{k=1}^{K}N_k(\mu_k-\mu)(\mu_k-\mu)^T\right\}}{\mathrm{tr}\left\{\sum_{k=1}^{K}\sum_{x\in C_k}(x-\mu_k)(x-\mu_k)^T\right\}}$$

[4] https://www.researchgate.net/publication/233096619_A_Dendrite_Method_for_

Индекс Калинского

- Используем свойства

$$\sum_i \operatorname{tr}\{\alpha_i A_i\} = \sum_i \alpha_i \operatorname{tr} A_i$$

$$\operatorname{tr}\{AB\} = \operatorname{tr}\{BA\}, \ \operatorname{tr} a = a \ \forall a \in \mathbb{R}$$

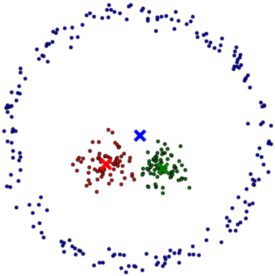
$$I = \frac{\operatorname{tr} B}{\operatorname{tr} W} = \frac{N-K}{K-1} \frac{\operatorname{tr}\left\{\sum_{k=1}^{K} N_k (\mu_k - \mu)(\mu_k - \mu)^T\right\}}{\operatorname{tr}\left\{\sum_{k=1}^{K} \sum_{x \in C_k} (x - \mu_k)(x - \mu_k)^T\right\}}$$

$$= \frac{N-K}{K-1} \frac{\sum_{k=1}^{K} N_k \operatorname{tr}\left\{(\mu_k - \mu)^T (\mu_k - \mu)\right\}}{\sum_{k=1}^{K} \sum_{x \in C_k} \operatorname{tr}\left\{(x - \mu_k)(x - \mu_k)^T\right\}} = \frac{N-K}{K-1} \frac{\sum_{k=1}^{K} N_k \|\mu_k - \mu\|^2}{\sum_{k=1}^{K} \sum_{x \in C_k} \|x - \mu_k\|^2}$$

- Измеряем отношение межкластерного к внутрикластерному разбросу.

Ограничение для невыпуклого кластера

- Сложность $O(ND)$, но поощряет выпуклые кластеры.
- Здесь качество Калинского будет казаться низким, как и индекс Дэвида-Болдуина:



- $\sum_{k=1}^{K} N_k \left\|\mu_k - \mu\right\|^2$ мало, а $\sum_{k=1}^{K} \sum_{x \in C_k} \left\|x - \mu_k\right\|^2$ велико
- Коэффициент силуэта будет вести себя лучше.

Алгоритм Monti consensus clustering[5]

- Генерируем H псевдовыборок $D_1, D_2, ...D_H$ из X, кластеризуем каждую.
- На D_h (x_i, x_j) кластеризуются как $\left(z_i^h, z_j^h\right)$, $h \in M\left(i, j\right)$.
  - $M\left(i,j\right) = \{h: \; x_i \in D_h \; \& \; x_j \in D_h\}$
- Определим матрицу консенсуса (consensus matrix)

$$M(i,j) = \frac{\sum_{h \in M(i,j)} \mathbb{I}\left[z_i^h = z_j^h\right]}{\left|M\left(i,j\right)\right|}, \quad M \in \mathbb{R}^{NxN}$$

- $M(i,j) \in \{0,1\}$ => у точек (i,j) устойчивая класт-ция.
- $M(i,j) \in (0,1)$ => у точек (i,j) неустойчивая класт-ция.
  - тем неустойчивее, чем ближе $M(i,j)$ к 0.5.

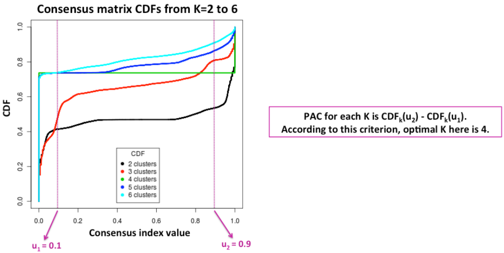
[5] https://en.wikipedia.org/wiki/Consensus_clustering

Алгоритм Monti consensus clustering

- Для каждого #кластеров K посчитаем пропорцию пар точек с неопределённой кластеризацией (proportion of ambiguous clustering, PAC)

$$PAC(K) = \frac{|\{(i,j) : i < j \ \& \ 0.1 \leq M(i,j) \leq 0.9\}|}{C_2^N}$$

- $PAC(K) \in [0,1]$ - мера неустойчивости кластеризации на K кластеров; $K^* = \arg\min_K PAC(K)$.



Содержание

Перекрестная таблица

- Пример перекрестной таблицы (contingency matrix):

	кластер 1	кластер 2	кластер 3
класс 1	5	2	0
класс 2	0	3	4

$\oplus$: Определяем разброс каждого класса по кластерам и разброс кластера по классам.

$\ominus$: Сложно анализировать для большого числа кластеров/классов. Не числовая метрика качества.

Перекрестная таблица

- Пример перекрестной таблицы (contingency matrix):

	кластер 1	кластер 2	кластер 3
класс 1	5	2	0
класс 2	0	3	4

$\oplus$: Определяем разброс каждого класса по кластерам и разброс кластера по классам.

$\ominus$: Сложно анализировать для большого числа кластеров/классов. Не числовая метрика качества.

- Числовая мера качества - Unsupervised Clustering Accuracy:
  - Π - всевозможные перенумеровки номеров кластеров

$$ACC(\boldsymbol{c}, \boldsymbol{z}) = \max_{\pi \in \Pi} \frac{1}{N} \sum_{n=1}^{N} \mathbb{I}\left[c_n = \pi\left(z_n\right)\right]$$

Матрица сочетаемости

- Матрица сочетаемости $\in \mathbb{R}^{2x2}$ вычисляет счётчики #пар (x_i, x_j).

	$z_i = z_j$	$z_i \neq z_j$
$y_i = y_j$	n_{11}	n_{12}
$y_i \neq y_j$	n_{21}	n_{22}

- Как понять по матрице качество кластеризации?

Матрица сочетаемости

- Матрица сочетаемости $\in \mathbb{R}^{2x2}$ вычисляет счётчики #пар (x_i, x_j).

	$z_i = z_j$	$z_i \neq z_j$
$y_i = y_j$	n_{11}	n_{12}
$y_i \neq y_j$	n_{21}	n_{22}

- Как понять по матрице качество кластеризации?

$\oplus$: Определяем сочетаемость разбиения по классам-кластерам.

$\ominus$: Не числовая метрика качества. Какую предложим?

Rand index

- Rand index - единая метрика по матрице сочетаемости.
- Пусть $y_1, ...y_N$ - истинная разметка. Обозначим[6]

$$n_{11} = |\{(x_i, x_j) : z_i = z_j \ \& \ y_i = y_j\}|$$

$$n_{22} = |\{(x_i, x_j) : z_i \neq z_j \ \& \ y_i \neq y_j\}|$$

$$\text{RandInd} = RI = \frac{n_{11} + n_{22}}{C_2^N} = \frac{n_{11} + n_{22}}{n_{11} + n_{12} + n_{21} + n_{22}} \in [0, 1]$$

- В чем недостаток?

[6] Это loss или score?

[7] Adjusted Rand Index - wikipedia.

[8] J-близость Жаккарда между множеством пар, у которых совпали классы и множеством пар, у которых совпали кластеры.

Rand index

- Rand index - единая метрика по матрице сочетаемости.
- Пусть $y_1, ...y_N$ - истинная разметка. Обозначим[6]

$$n_{11} = |\{(x_i, x_j) : z_i = z_j \ \& \ y_i = y_j\}|$$

$$n_{22} = |\{(x_i, x_j) : z_i \neq z_j \ \& \ y_i \neq y_j\}|$$

$$\text{RandInd} = RI = \frac{n_{11} + n_{22}}{C_2^N} = \frac{n_{11} + n_{22}}{n_{11} + n_{12} + n_{21} + n_{22}} \in [0, 1]$$

- В чем недостаток?$\uparrow RI$ с $\uparrow$#кластеров. Лучше[7,8]

$$\text{AdjRandInd} = \frac{RI - \mathbb{E}\{RI\}}{\max(RI) - \mathbb{E}\{RI\}} \text{ или } \mathrm{J} = \frac{n_{11}}{n_{11} + n_{12} + n_{21}}$$

---

[6] Это loss или score?

[7] Adjusted Rand Index - wikipedia.

[8] J-близость Жаккарда между множеством пар, у которых совпали классы и множеством пар, у которых совпали кластеры.

Гомогенность[9]

- Обозначим N=#объектов, n_k=#объектов в кластере k, m_c=#объектов в классе c, n_{ck}=#объектов класса c в кластере k.

$$H_{class} = -\sum_{c=1}^{C} \frac{m_c}{N} \log \frac{m_c}{N}$$

$$H_{clust} = -\sum_{k=1}^{K} \frac{n_k}{N} \log \frac{n_k}{N}$$

$$H_{class|clust} = -\sum_{k=1}^{K} \frac{n_k}{N} \sum_{c=1}^{C} \frac{n_{ck}}{n_k} \log \frac{n_{ck}}{n_k}$$

$H_{class|clust} = 0$ при полном объяснении, $H_{class|clust} = 1$ нет связи

[9] https://aclanthology.org/D07-1043.pdf

Гомогенность

$$\text{Homogeneity} = 1 - \frac{H\left(class|clust\right)}{H\left(class\right)}$$

- Гомогенность показывает долю информации о классах, объясненной кластеризацией.
    - 1: в кластерах представители только 1 класса
    - 0: в кластерах распределение классов=априорному распределению
- Какой недостаток?

Гомогенность

$$\text{Homogeneity} = 1 - \frac{H\left(class|clust\right)}{H\left(class\right)}$$

- Гомогенность показывает долю информации о классах, объясненной кластеризацией.
    - 1: в кластерах представители только 1 класса
    - 0: в кластерах распределение классов=априорному распределению
- Какой недостаток? Гомогенность поощряет ↑#кластеров
    - =1, когда каждый объект - в своём кластере

Полнота[10]

- Нужна доп. мера полноты (насколько объекты одного класса оказываются в одном кластере)

$$\text{Completeness} = 1 - \frac{H\left(clust|class\right)}{H\left(clust\right)}$$

- Полнота =1, если класс полностью определяет кластер (все объекты кластера-в одном классе)
- Какой недостаток?

[10] https://aclanthology.org/D07-1043.pdf

Полнота[10]

- Нужна доп. мера полноты (насколько объекты одного класса оказываются в одном кластере)

$$\text{Completeness} = 1 - \frac{H\left(clust|class\right)}{H\left(clust\right)}$$

- Полнота =1, если класс полностью определяет кластер (все объекты кластера-в одном классе)
- Какой недостаток? Полнота поощряет $\downarrow$#кластеров
  - =1, когда все объекты в одном кластере

[10] https://aclanthology.org/D07-1043.pdf

V-мера[11]

- V-мера - среднее гармоническое от гомогенности и полноты.
$$V=\frac{1}{\frac{1}{2}\frac{1}{\text{Homogeniety}}+\frac{1}{2}\frac{1}{\text{Completeness}}}$$
- Взвешенный учёт гомогенности и полноты:
$$V_\beta=\frac{1}{\left(\frac{\beta}{1+\beta}\right)\frac{1}{\text{Homogeniety}}+\frac{1}{1+\beta}\frac{1}{\text{Completeness}}}$$
- $V=V_\beta$ при $\beta=1$.

[11] https://aclanthology.org/D07-1043.pdf

Нормализованная взаимная информация

- Взаимная информация - степень связи сл. вел. X, Y:

$$MI(X,Y) = KL(P(X,Y)||P(X)P(Y))$$
$$= \sum_{x\in dom(X)} \sum_{y\in dom(Y)} P(x,y)\log\frac{P(x,y)}{P(x)P(y)}$$

$$MI(X,Y) = H(Y) - H(Y|X) = H(X) - H(X|Y)$$

- Нормализованная взаимная информация $(NMI \in [0,1])$[12] - др. вариант агрегации полноты и гомогенности:

$$NMI(clust, class) = \frac{MI(clust, class)}{\max\{H_{clust}, H_{class}\}}$$
$$= \frac{H(clust) - H(clust|class)}{\max\{H_{clust}, H_{class}\}} = \frac{H(class) - H(class|clust)}{\max\{H_{clust}, H_{class}\}}$$

[12] Это loss или score?

Заключение

- Оценки, не использующие разметку:
    - размеры кластеров vs. межкластерные расстояния
- Оценки, использующие разметку:
    - сопоставление кластеров с истинными метками
        - важна инвариантность к перенумеровке кластеров