

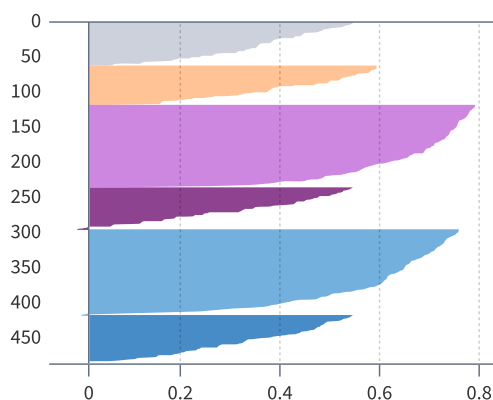
Силуэт кластера (Cluster silhouette)

Синонимы: Диаграмма силуэта, Silhouette diagram

Разделы: [Визуализация](#)

Силуэт кластера — метод графического представления результатов кластеризации, с помощью которого можно визуально оценить качество построенной кластерной модели.

В основе идеи метода лежит вычисление коэффициентов кластерных силуэтов. На диаграмме для каждого объекта коэффициент силуэта отображается прямоугольником соответствующей длины. Прямоугольники группируются по кластерам (которые обычно выделяются цветом) и в каждом кластере дополнительно ранжируются в порядке убывания.

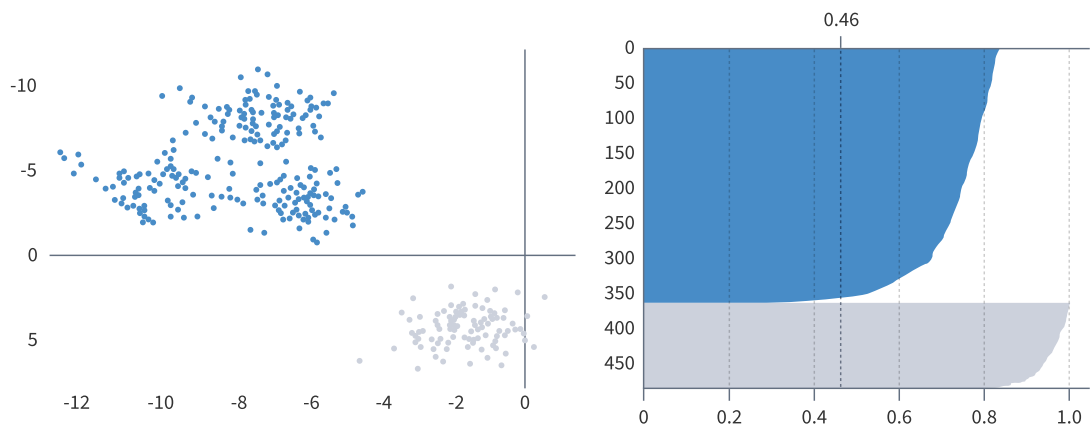


Таким образом, на диаграмме становится виден «силуэт» каждого кластера, откуда и название метода. По форме силуэтов аналитик оперативно может оценить качество кластеризации. Чем форма силуэтов ближе к прямоугольной, а площадь (средний коэффициент силуэта) ближе к 1, тем лучше кластеризация. Внутри силуэта каждого кластера объекты расположены в порядке убывания их коэффициента силуэта, поэтому легко увидеть, какие именно объекты лучше соответствуют кластеру, а какие хуже.

Напротив, чем больше в кластере объектов с низким коэффициентом силуэта, которые порождают «узкие» силуэты, тем хуже кластеризация.

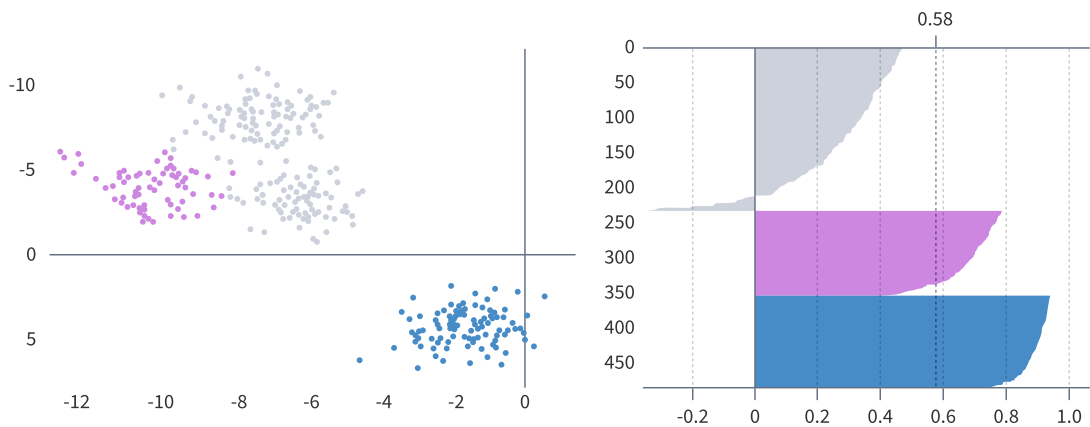
Таким образом, диаграммы силуэтов и средние значения коэффициентов могут использоваться для определения естественного числа кластеров в наборе данных. Поясним сказанное с помощью рисунков.

На следующем рисунке представлено распределение точек данных (слева) в 2-мерном пространстве признаков и диаграмма силуэтов (справа) для случая 2-х кластеров.

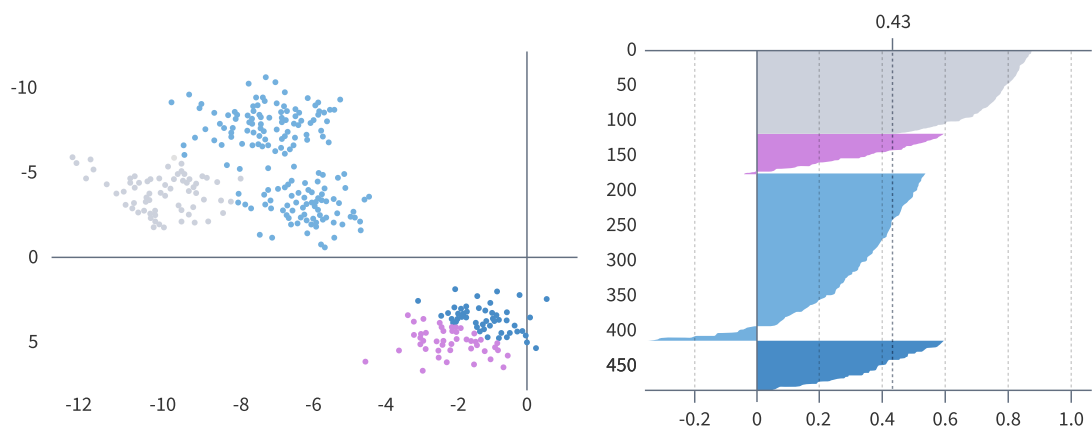


Ширина силуэтов обоих кластеров превышает среднее значение коэффициента силуэта, равное 0.46. Это говорит о том, что модель, содержащая два кластера, хорошо соответствует естественной группировке данных.

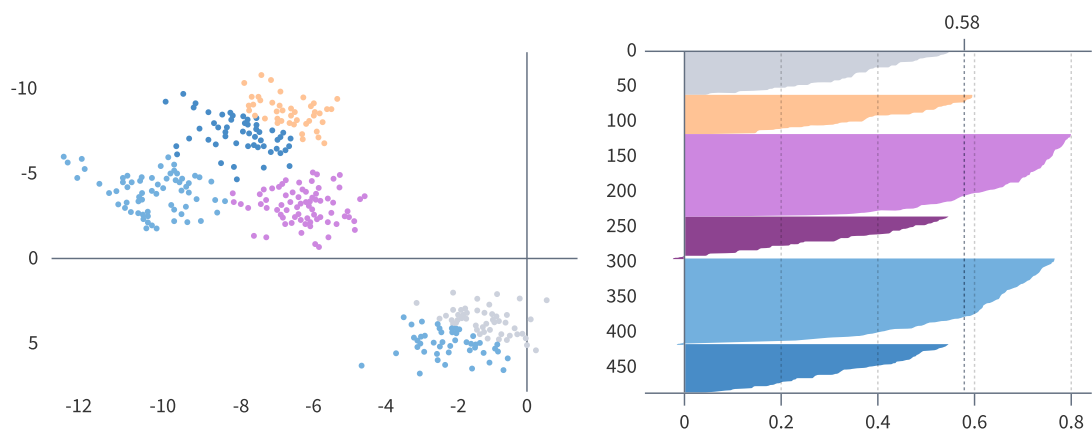
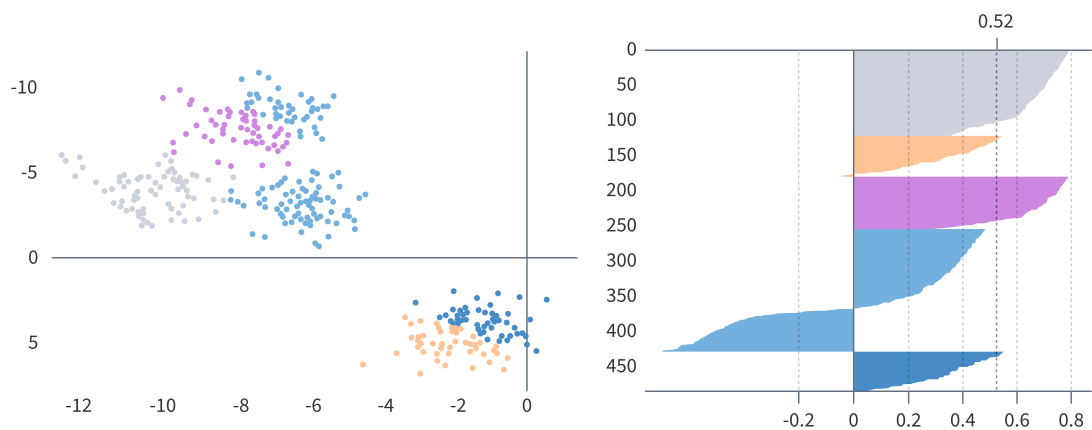
Для структуры, содержащей 3 кластера несложно увидеть, что 0-й кластер имеет силуэт, в котором коэффициент силуэта ни для одного объекта не превышает среднее значение коэффициента силуэта, равное 0.58. Это говорит о том, что в данном случае соответствие модели естественной группировке несколько хуже, чем для случая 2-х кластеров.



На следующем рисунке представлены силуэты для 4-кластерной модели. Хорошо видно, что все кластерные силуэты имеют ширину, превышающие среднее значение 0.43, что говорит о хорошем соответствии кластерной структуры исходным данным.



И, наконец, 5 и 6-кластерные модели не являются оптимальными, поскольку в их диаграммах силуэтов содержатся три «узких» кластера, ширина силуэтов которых не превышает среднее значение индекса силуэта 0.52 и 0.58 соответственно.



Подбирая параметры кластеризации (число кластеров, веса признаков специфичных для кластеров и т.д.) и отслеживая результаты по диаграмме силуэтов, удобно сравнивать полученные кластерные модели и выбирать те, которые наилучшим образом соответствуют задаче анализа данных.

Например, несложно увидеть, что в рассмотренном примере число кластеров, при котором модель будет наилучшим образом соответствовать данным, равно двум. Это видно по тому, что силуэты кластеров на диаграмме наиболее широкие. При увеличении числа кластеров на диаграмме появляются узкие силуэты, а это говорит о том, что число кластеров начало превышать число естественных групп в данных.

Научиться оценивать качество кластеризации и выбирать оптимальное число кластеров можно в рамках воркшопа «K-means и кластерные силуэты».