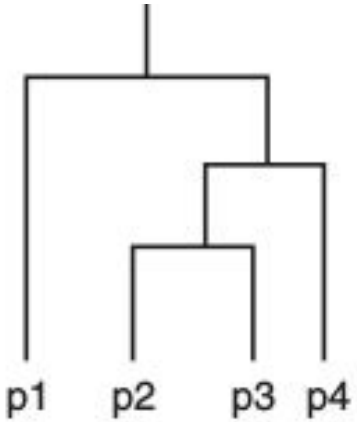
Кластеризация

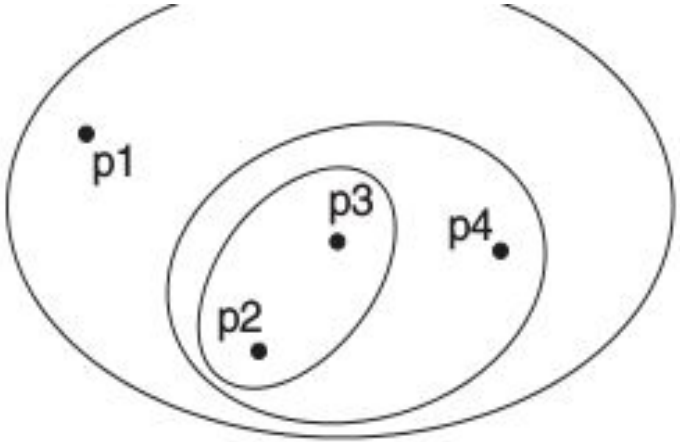
Иерархическая кластеризация

- Agglomerative (Агломеративная)
- Divisive (Дивизионная)

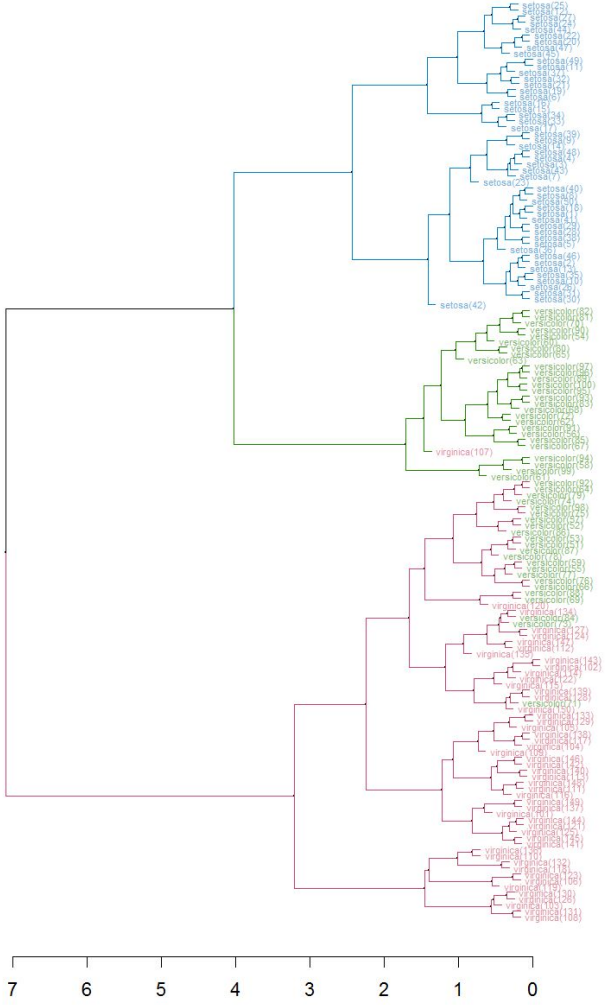
Графическое представление



(a) Dendrogram.



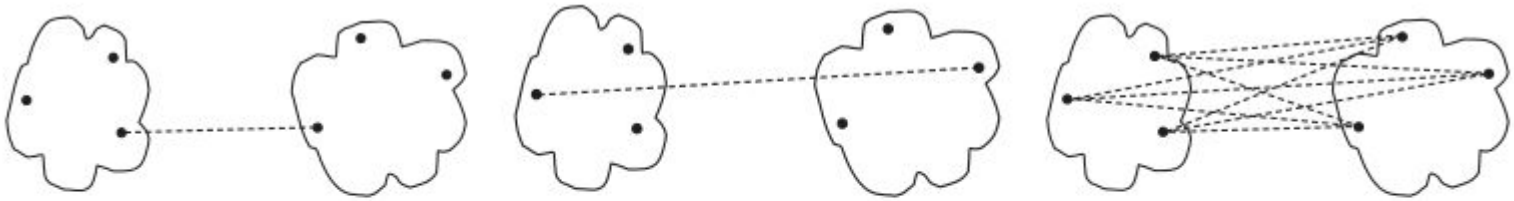
(b) Nested cluster diagram.


7
6
5
4
3
2
1
0

Basic Agglomerative Hierarchical Clustering Algorithm

1. Compute the proximity matrix, if necessary.
2. **repeat**
3. Merge the closest two clusters.
4. Update the proximity matrix to reflect the proximity between the new cluster and the original clusters.
5. **until** Only one cluster remains

Определение близости между кластерами



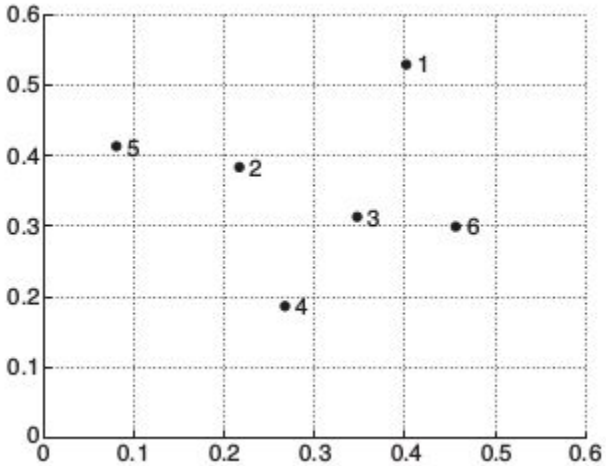
(a) MIN (single link.) (b) MAX (complete link.) (c) Group average.

Определение близости между кластерами

$$R(U \cup V, S) = \alpha_U R(U, S) + \alpha_V R(V, S) + \\ + \beta R(U, V) + \gamma |R(U, S) - R(V, S)|$$

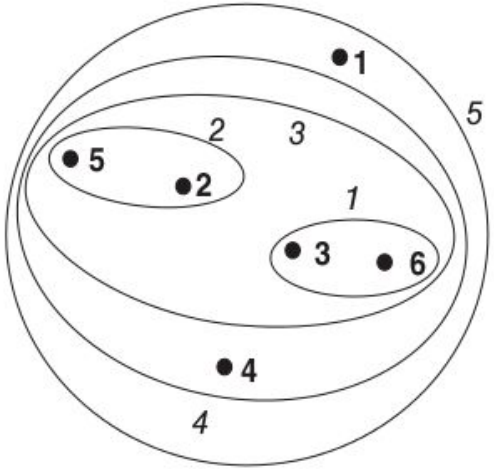
Пример

	p1	p2	p3	p4	p5	p6
p1	0.00	0.24	0.22	0.37	0.34	0.23
p2	0.24	0.00	0.15	0.20	0.14	0.25
p3	0.22	0.15	0.00	0.15	0.28	0.11
p4	0.37	0.20	0.15	0.00	0.29	0.22
p5	0.34	0.14	0.28	0.29	0.00	0.39
p6	0.23	0.25	0.11	0.22	0.39	0.00

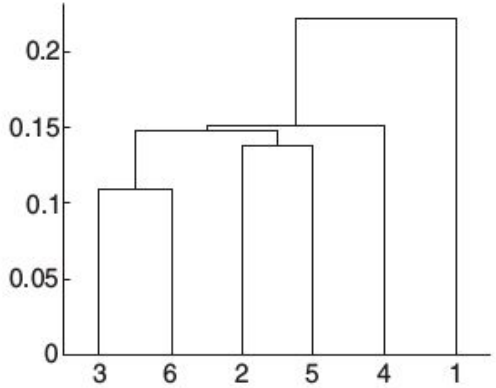


Point	x Coordinate	y Coordinate
p1	0.40	0.53
p2	0.22	0.38
p3	0.35	0.32
p4	0.26	0.19
p5	0.08	0.41
p6	0.45	0.30

MIN

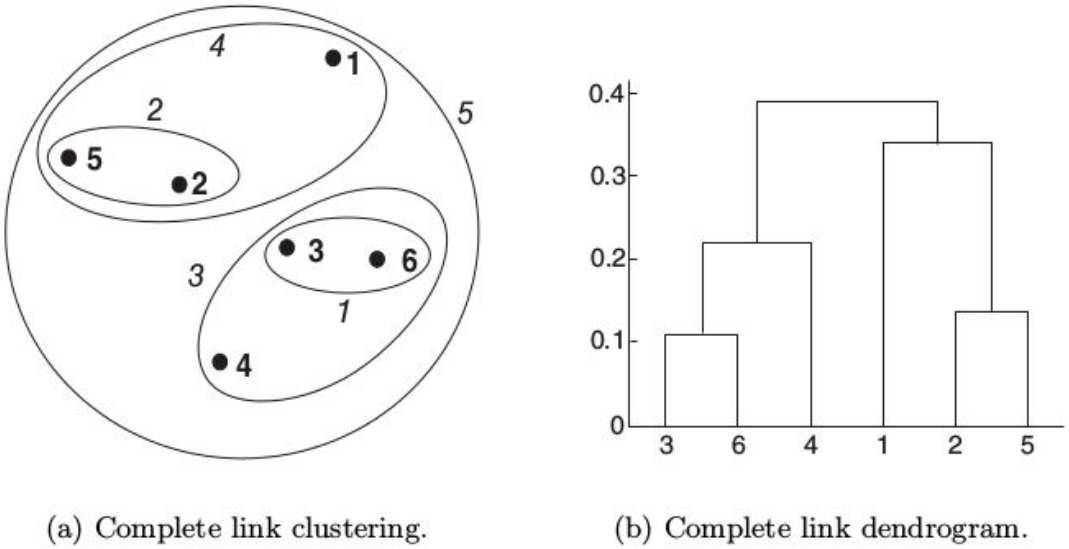


(a) Single link clustering.



(b) Single link dendrogram.

MAX

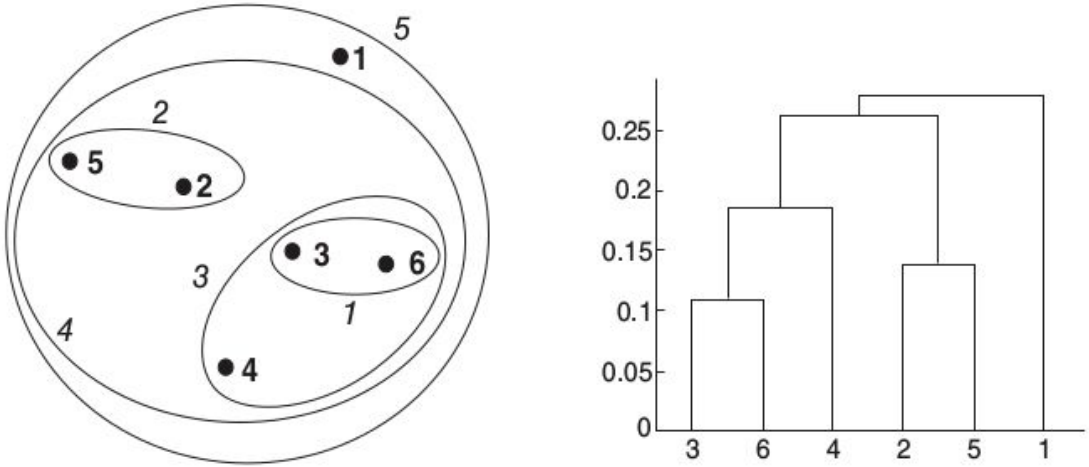


(a) Complete link clustering.

(b) Complete link dendrogram.

Figure 8.17. Complete link clustering of the six points shown in Figure 8.15.

Group Average



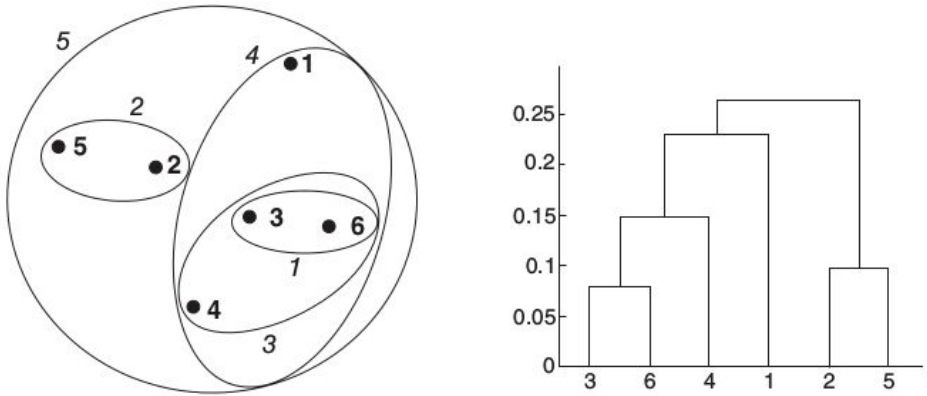
(a) Group average clustering.

(b) Group average dendrogram.

Figure 8.18. Group average clustering of the six points shown in Figure 8.15.

Table 8.5. Table of Lance-Williams coefficients for common hierarchical clustering approaches.

Clustering Method	α_A	α_B	β	γ
Single Link	$1/2$	$1/2$	0	$-1/2$
Complete Link	$1/2$	$1/2$	0	$1/2$
Group Average	$\frac{m_A}{m_A+m_B}$	$\frac{m_B}{m_A+m_B}$	0	0
Centroid	$\frac{m_A}{m_A+m_B}$	$\frac{m_B}{m_A+m_B}$	$\frac{-m_A m_B}{(m_A+m_B)^2}$	0
Ward's	$\frac{m_A+m_Q}{m_A+m_B+m_Q}$	$\frac{m_B+m_Q}{m_A+m_B+m_Q}$	$\frac{-m_Q}{m_A+m_B+m_Q}$	0

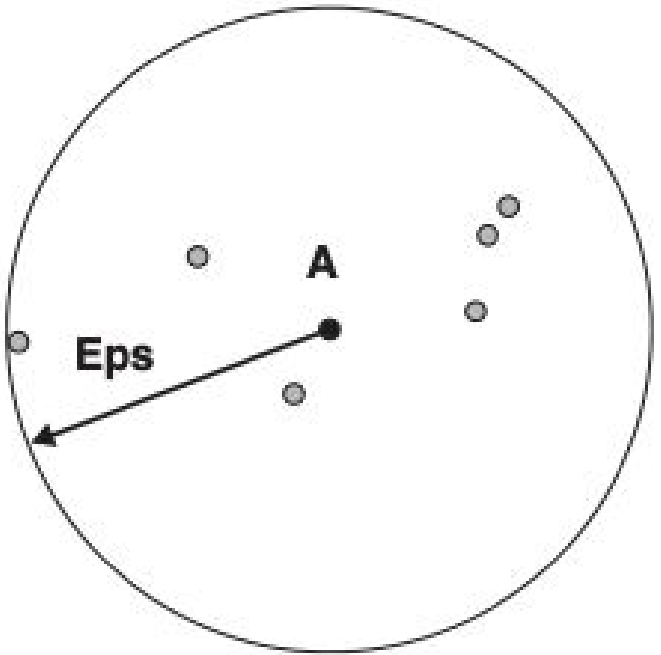


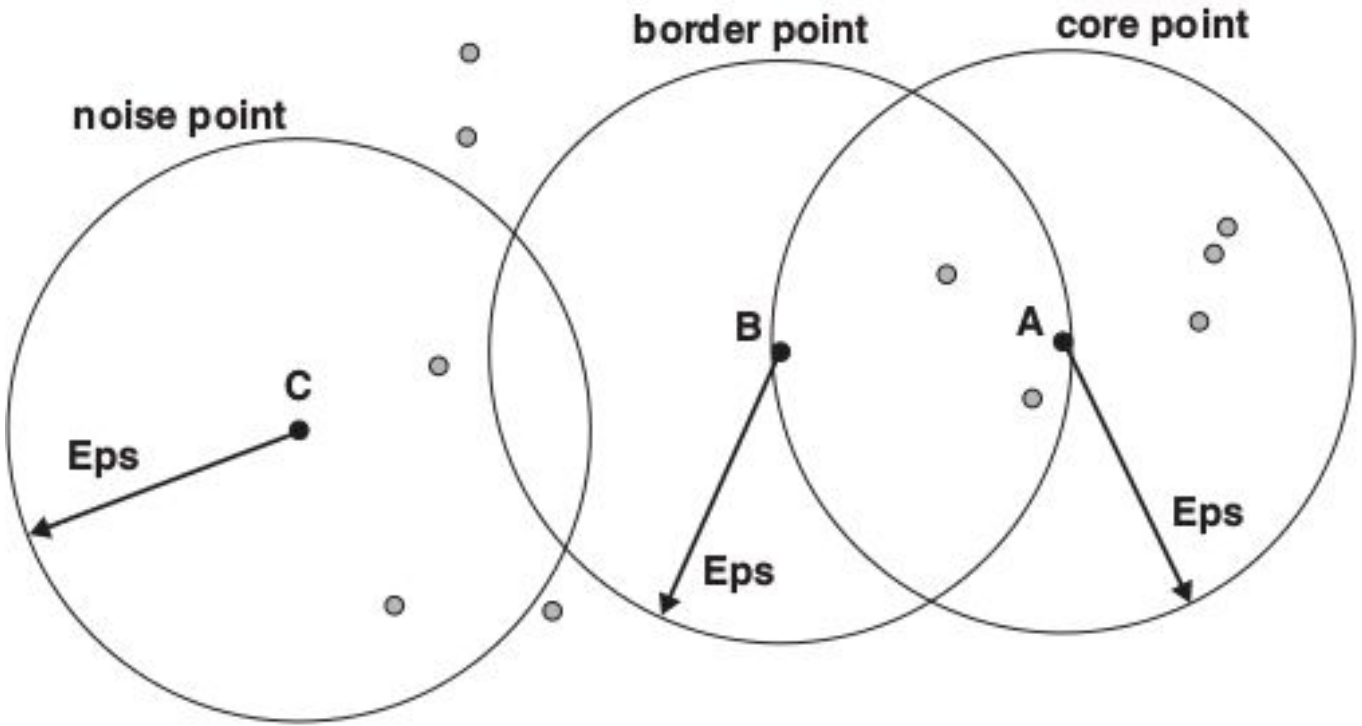
(a) Ward's clustering. (b) Ward's dendrogram.

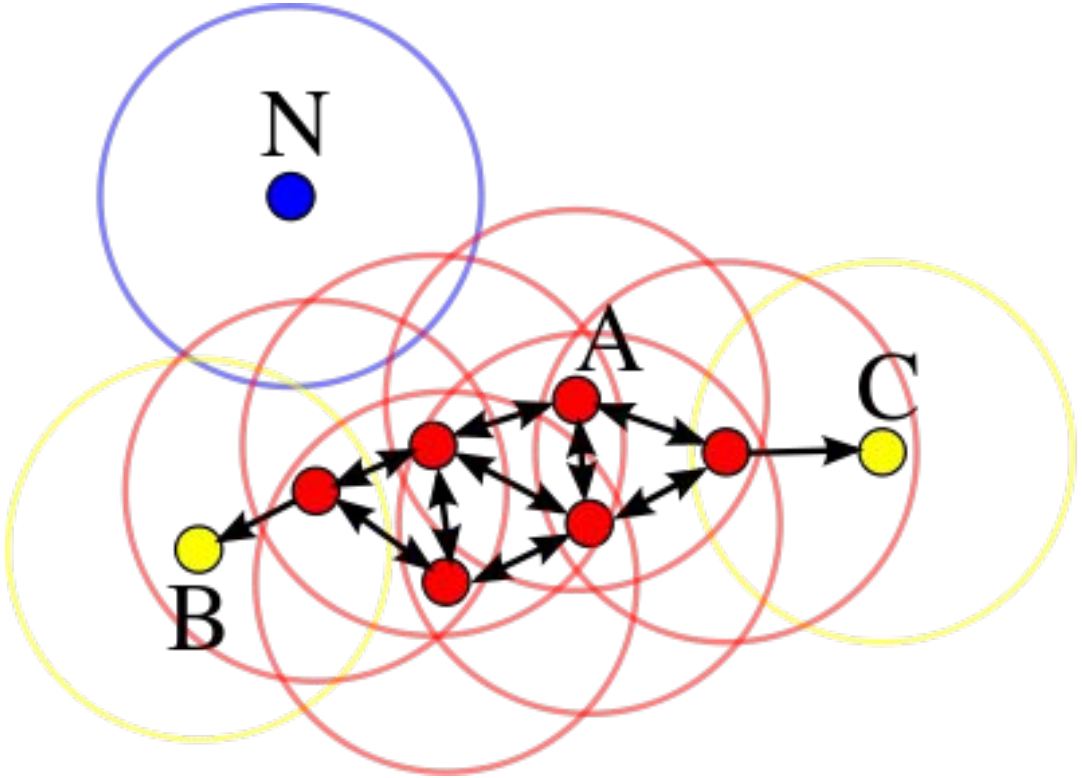
Figure 8.19. Ward's clustering of the six points shown in Figure 8.15.

DBSCAN

Кластеризация на основе плотности точек




border point
core point
noise point
B
A
C
Eps
Eps
Eps


N
A
C
B

The DBSCAN Algorithm

1. Label all points as core, border, or noise points.
2. Eliminate noise points.
3. Put an edge between all core points that are within *Eps* of each other.
4. Make each group of connected core points into a separate cluster.
5. Assign each border point to one of the clusters of its associated core points.

Определение параметров

Figure 8.22. Sample data.

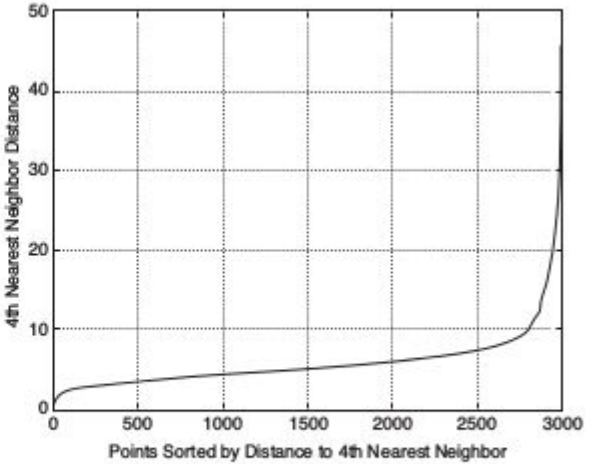
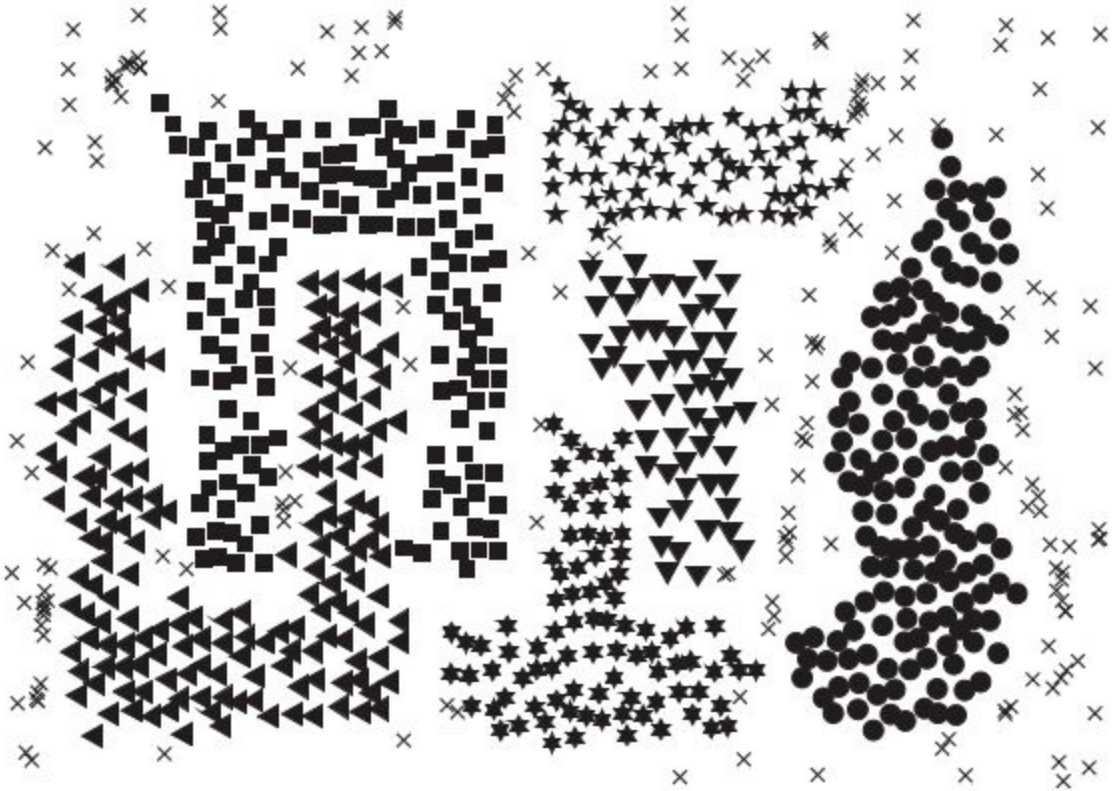
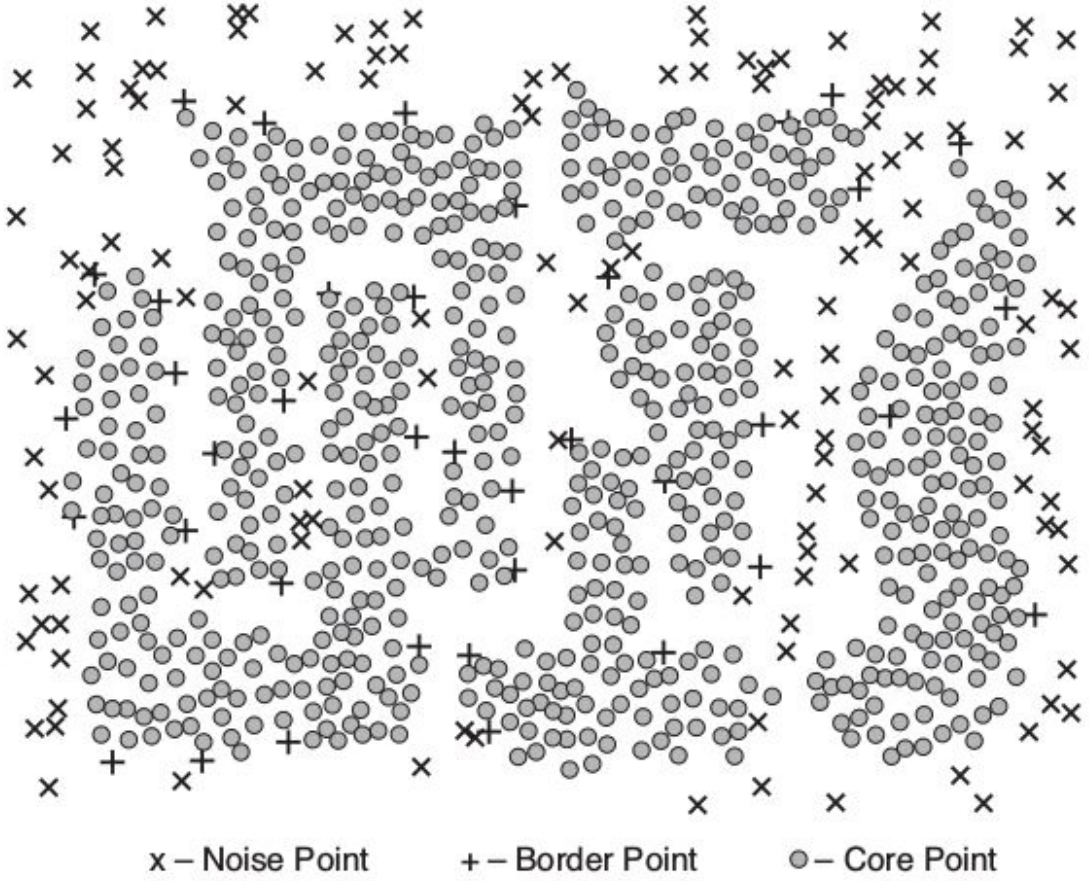


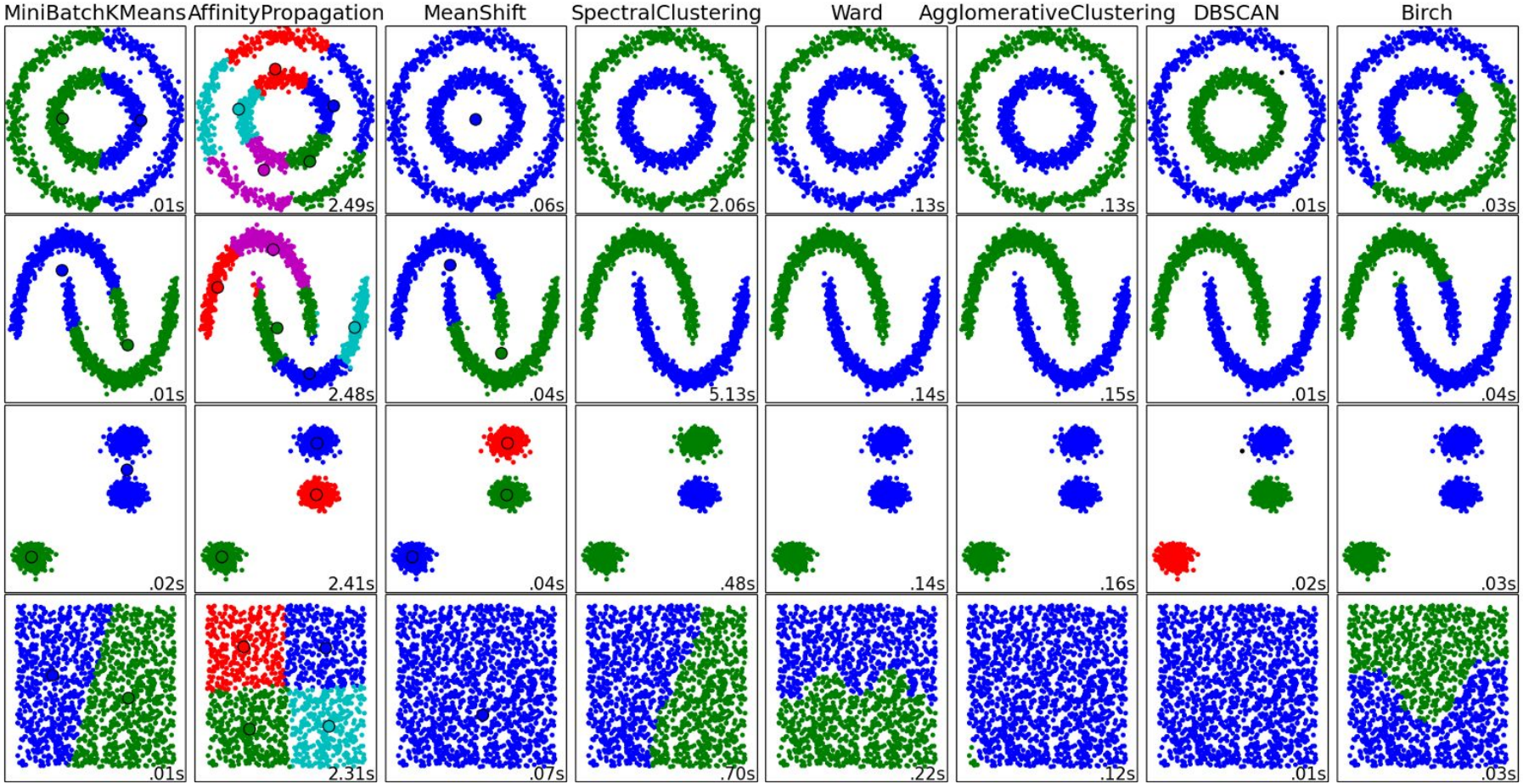
Figure 8.23. K-dist plot for sample data.



(a) Clusters found by DBSCAN.



(b) Core, border, and noise points.


MiniBatchKMeans
AffinityPropagation
MeanShift
SpectralClustering
Ward
AgglomerativeClustering
DBSCAN
Birch
.01s
2.49s
.06s
2.06s
.13s
.13s
.01s
.03s
.01s
2.48s
.04s
5.13s
.14s
.15s
.01s
.04s
.02s
2.41s
.04s
.48s
.14s
.16s
.02s
.03s
.01s
2.31s
.07s
.70s
.22s
.12s
.01s
.03s

Решение задачи кластеризации

Среднее внутрикластерное расстояние

$$F_0 = \frac{\sum_{i<j} [y_i = y_j] \rho(x_i, x_j)}{\sum_{i<j} [y_i = y_j]} \to \min$$

$$F_0' = \sum_{y \in Y} \frac{1}{|K_y|} \sum_{i: y_i = y} \rho^2(x_i, \mu_y) \to \min$$

Решение задачи кластеризации

Среднее межкластерное расстояние

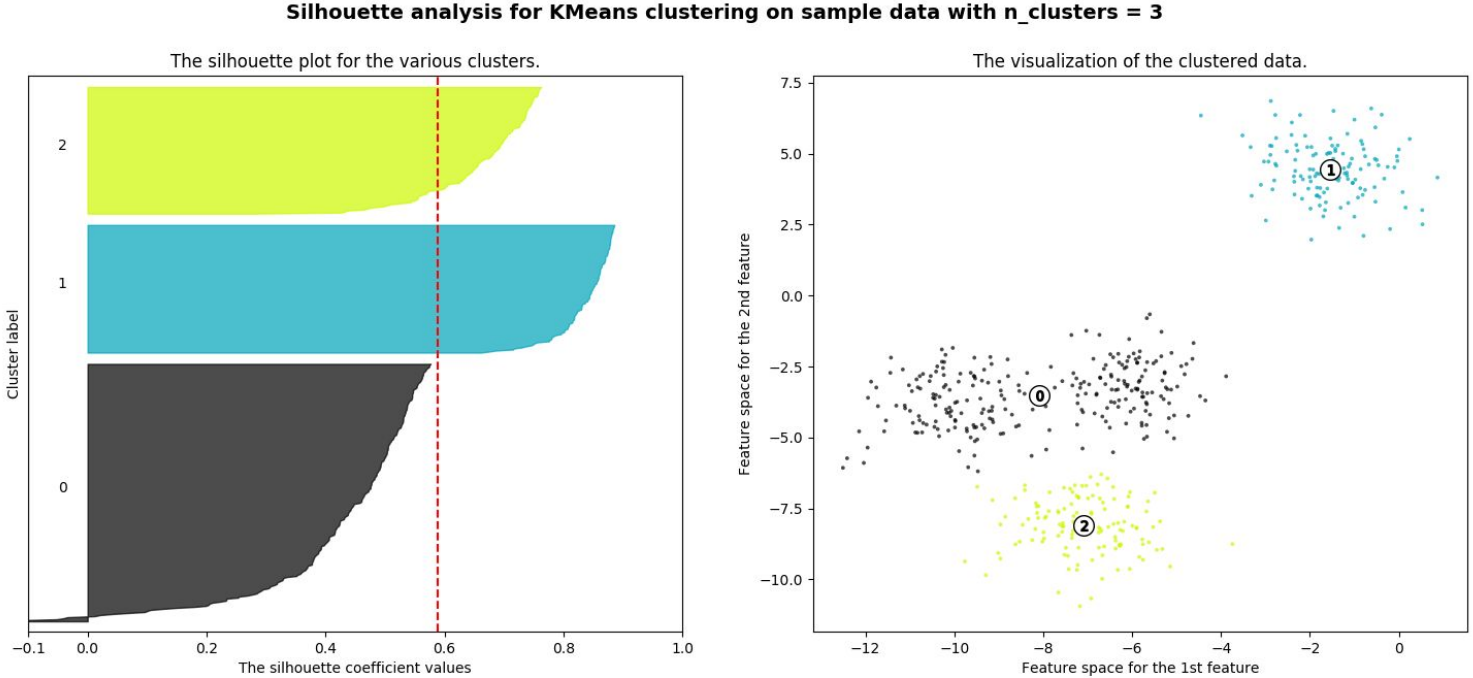
$$F_1 = \frac{\sum_{i<j} [y_i \neq y_j] \rho(x_i, x_j)}{\sum_{i<j} [y_i \neq y_j]} \to \max$$

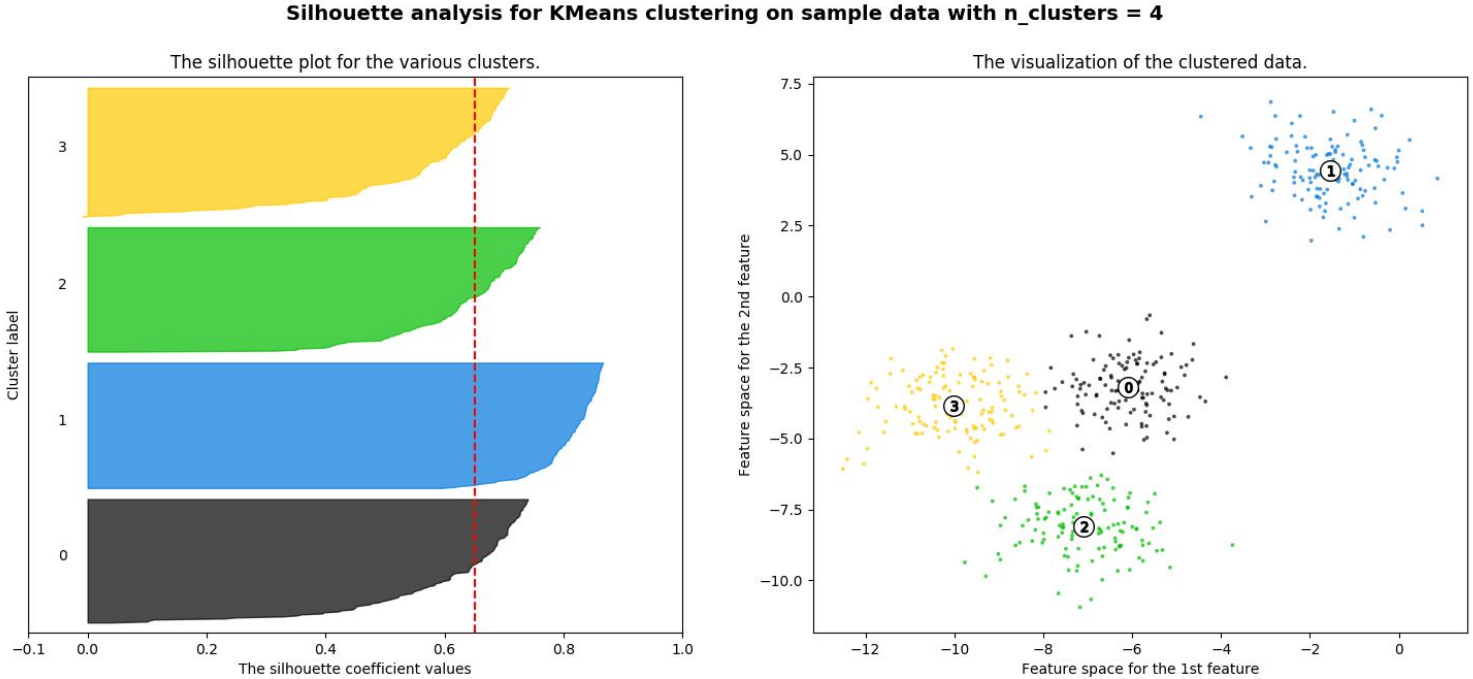
$$F_1' = \sum_{y \in Y} \rho^2(\mu_y, \mu) \to \max$$

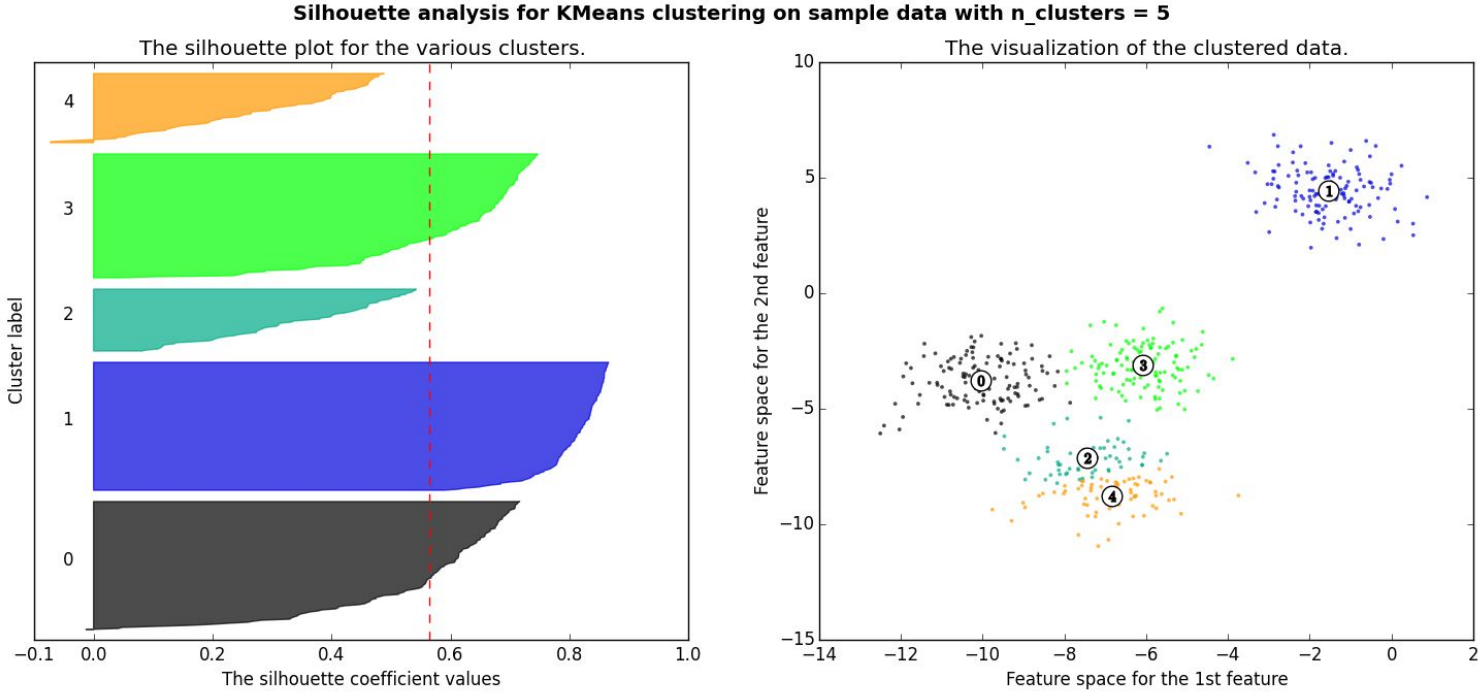
Коэффициент силуэта

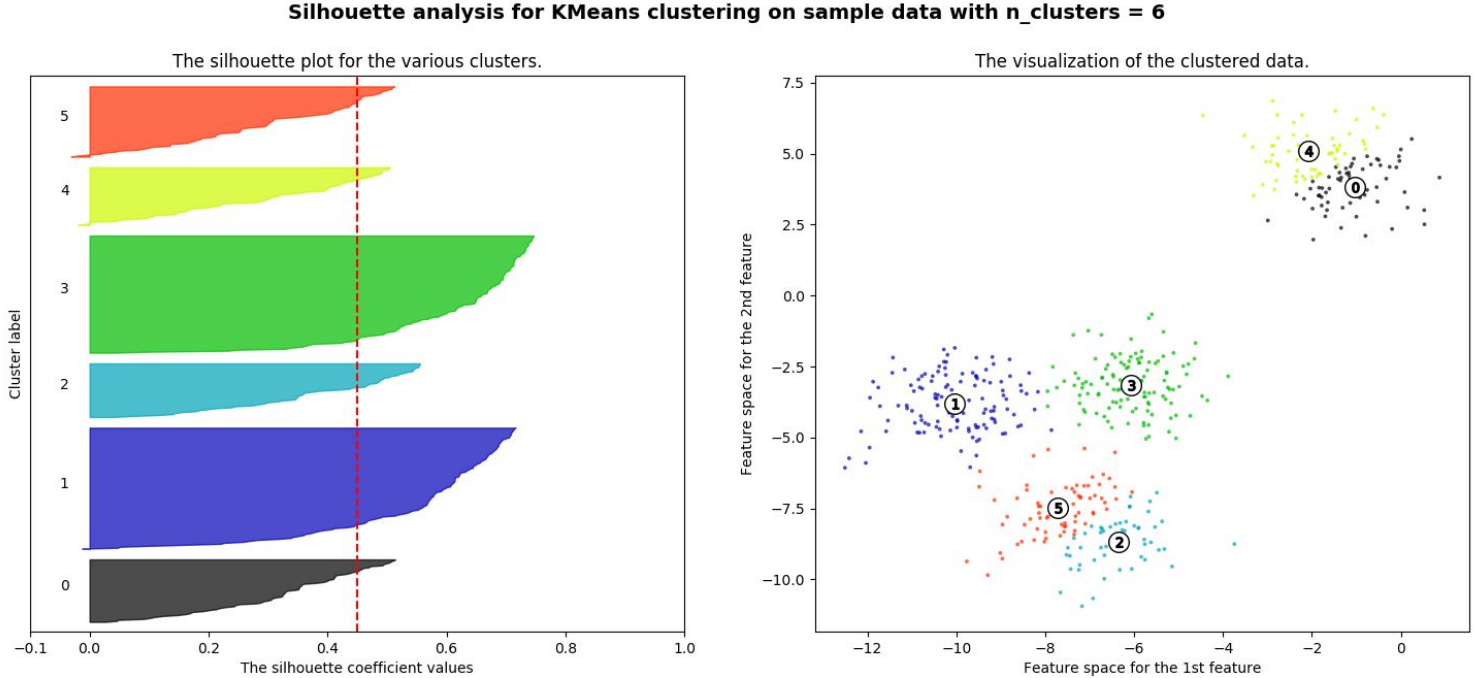
a - среднее расстояние от объекта до всех других объектов из этого кластера
b - среднее расстояние от объекта до всех объектов из ближайшего другого кластера

$$s = \frac{b - a}{\max(a, b)}$$


Silhouette analysis for KMeans clustering on sample data with n_clusters = 3
The silhouette plot for the various clusters.
Cluster label
2
1
0
−0.1
0.0
0.2
0.4
0.6
0.8
1.0
The silhouette coefficient values
The visualization of the clustered data.
Feature space for the 2nd feature
7.5
5.0
2.5
0.0
−2.5
−5.0
−7.5
−10.0
−12
−10
−8
−6
−4
−2
0
Feature space for the 1st feature
0
1
2


Silhouette analysis for KMeans clustering on sample data with n_clusters = 4
The silhouette plot for the various clusters.
Cluster label
3
2
1
0
−0.1
0.0
0.2
0.4
0.6
0.8
1.0
The silhouette coefficient values
The visualization of the clustered data.
Feature space for the 2nd feature
7.5
5.0
2.5
0.0
−2.5
−5.0
−7.5
−10.0
−12
−10
−8
−6
−4
−2
0
Feature space for the 1st feature
1
0
3
2


Silhouette analysis for KMeans clustering on sample data with n_clusters = 5
The silhouette plot for the various clusters.
Cluster label
4
3
2
1
0
−0.1
0.0
0.2
0.4
0.6
0.8
1.0
The silhouette coefficient values
The visualization of the clustered data.
Feature space for the 2nd feature
10
5
0
−5
−10
−15
−14
−12
−10
−8
−6
−4
−2
0
2
Feature space for the 1st feature
0
1
2
3
4


Silhouette analysis for KMeans clustering on sample data with n_clusters = 6
The silhouette plot for the various clusters.
Cluster label
The silhouette coefficient values
The visualization of the clustered data.
Feature space for the 2nd feature
Feature space for the 1st feature

Описание изображений

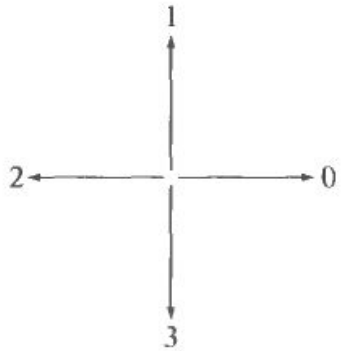
Задачи

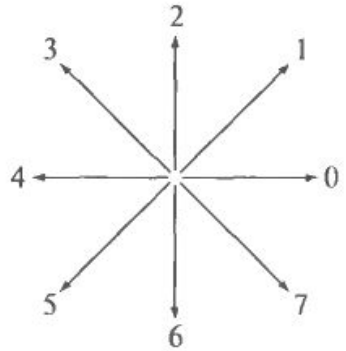
- Представление
- Описание

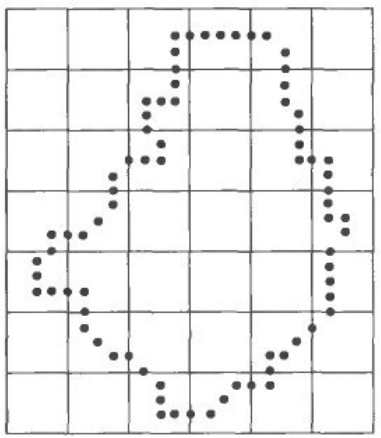
Представление

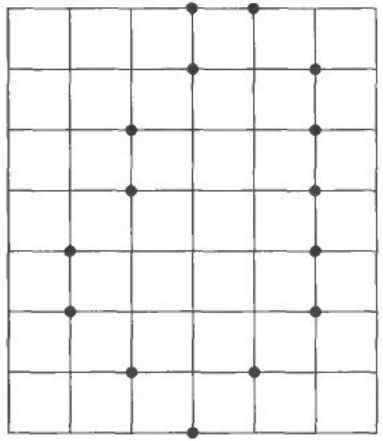
Цепные коды (Код Фримена)

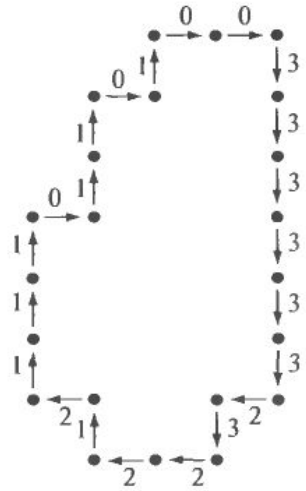
Граница - последовательность соединенных отрезков, для которых указаны длина и направление.

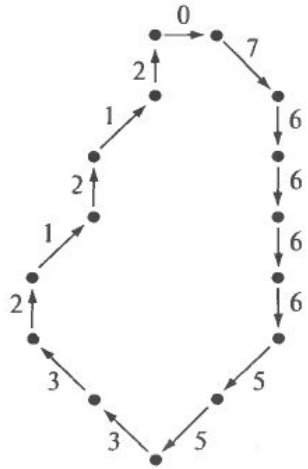








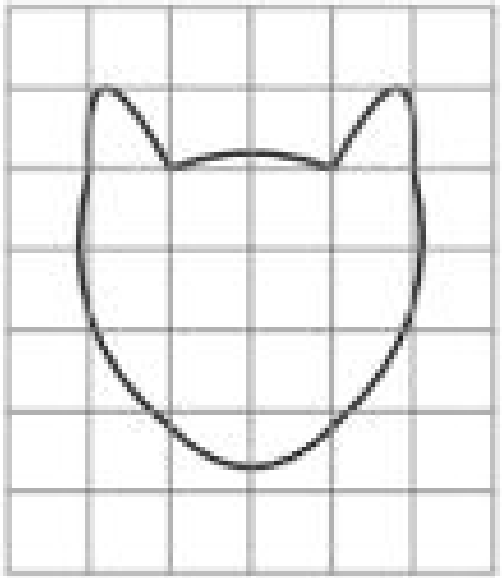

0 0
1 0
3
1 3
1 3
0 1 3
1 3
1 3
1 3
1 3
2 1 3 2
2 2

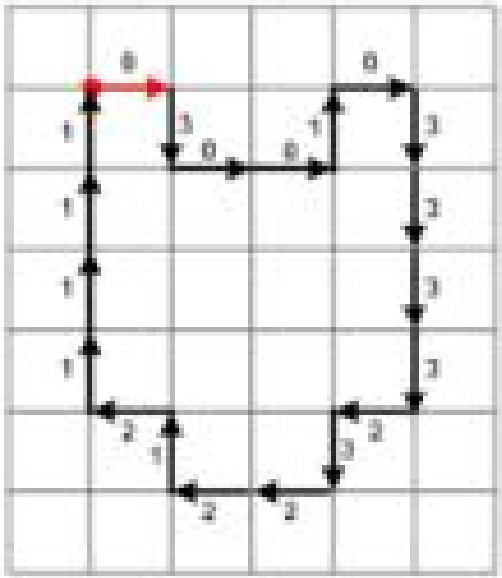

0
7
2 6
1 6
2 6
1 6
2 6
3 5
3 5

Инвариантность

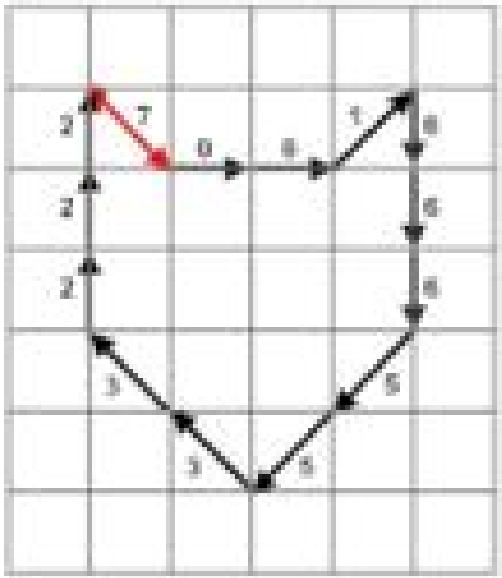
- Относительно начальной точки
- Относительно поворота
- Относительно размера

Пример





А



Б

Аппроксимация ломаной линией

Точная аппроксимация - число отрезков ломаной равно числу точек границы

Ломаные минимальной длины

Если каждый элемент заключает в себе единственную точку границы, то величина отклонения фактической границы от её приближения лентой внутри любого элемента не превышает $\sqrt{2}d$

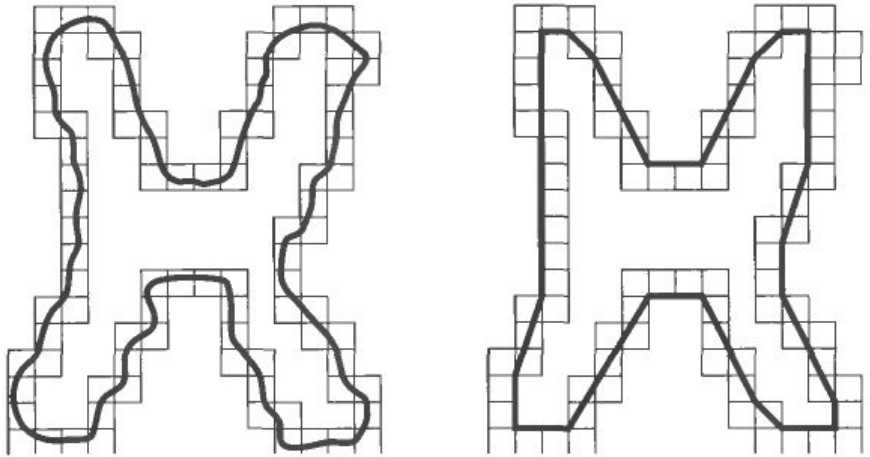


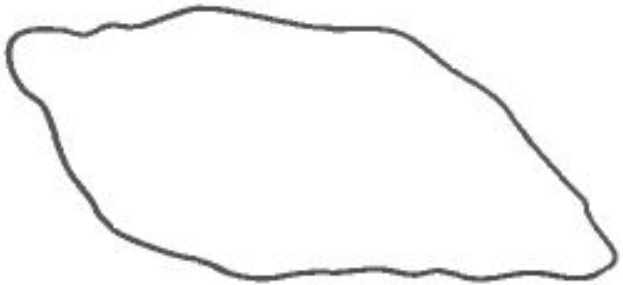
Рис. 11.3. (а) Граница объекта, заключенная внутри цепочки элементов. (б) Ломаная минимальной длины.

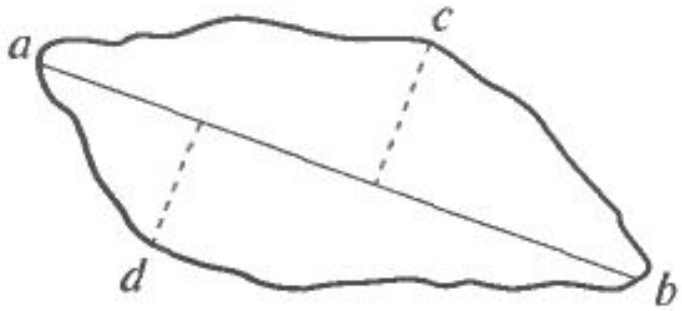
Методы слияния

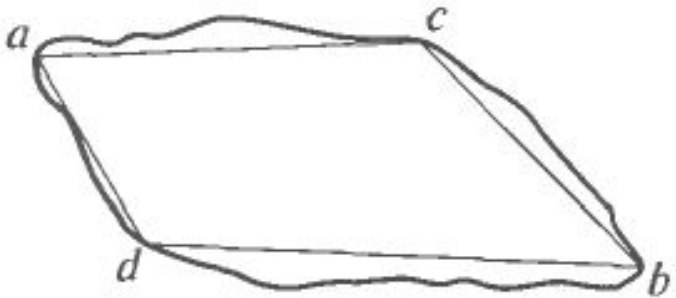
Основаны на применении к задаче кусочно-линейной аппроксимации критерия средней ошибки или критерия другого вида

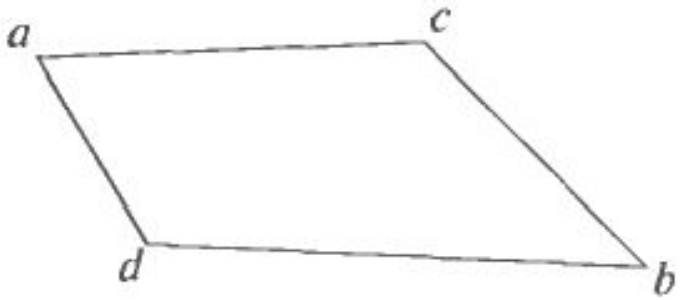
Методы разбиения

Отрезок последовательно разбивается на 2 части, пока не начнет удовлетворять заданному критерию



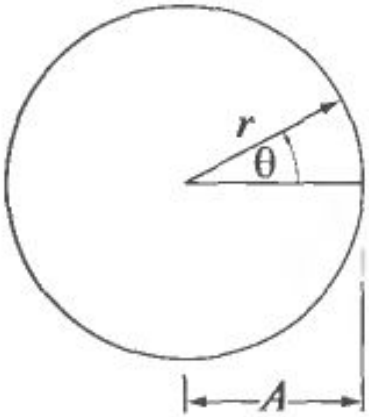

a
c
d
b

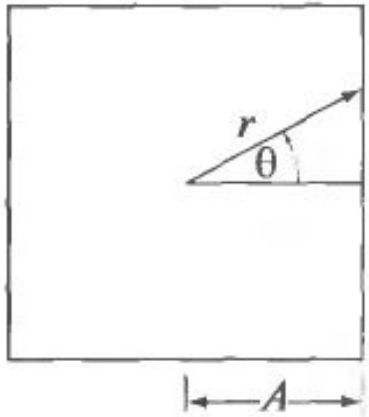

a
c
d
b

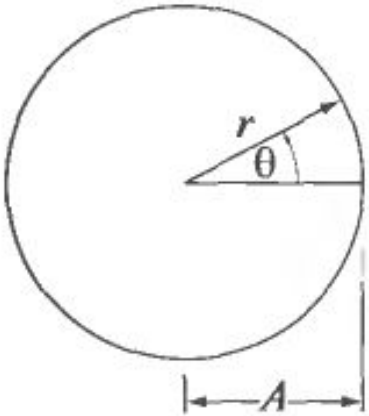

a
c
d
b

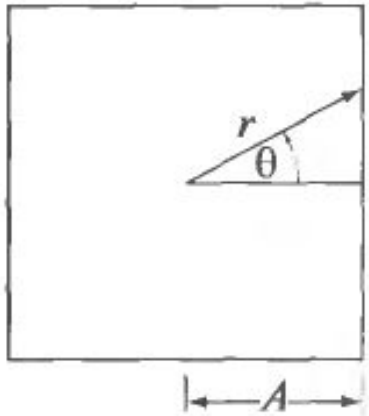
Сигнатуры

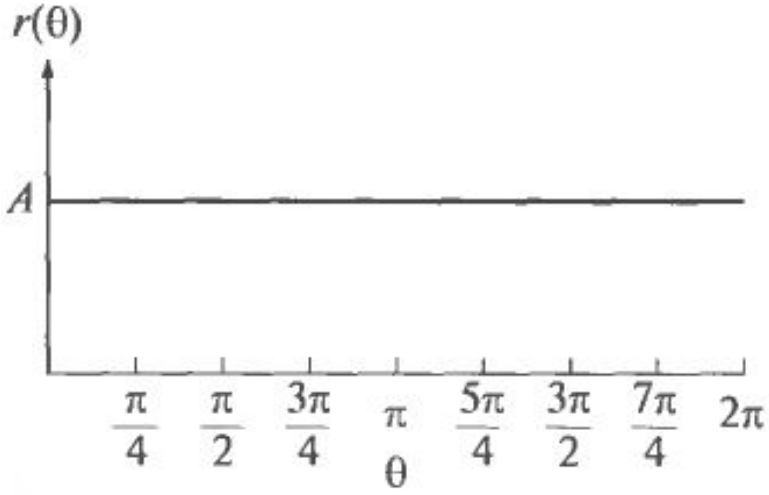
Описание объекта с помощью одномерной функции

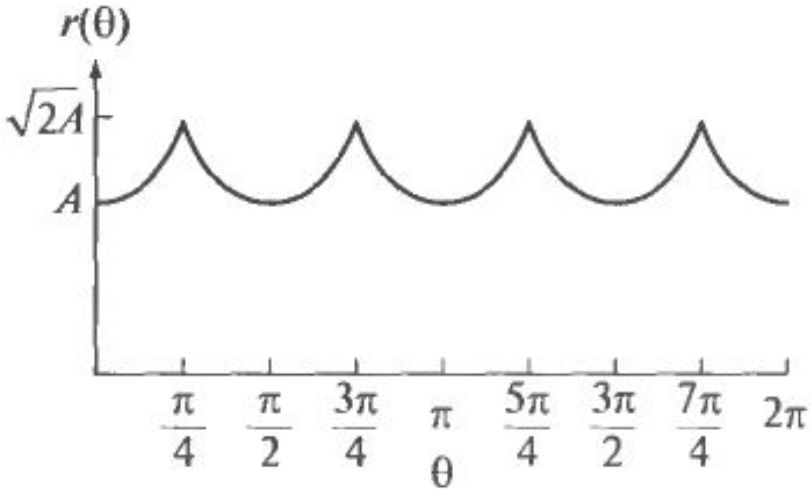

r
θ
A


r
θ
A


r
θ
A


r
θ
A


r(θ)
A
π/4
π/2
3π/4
π
θ
5π/4
3π/2
7π/4
2π


r(θ)
√2A
A
π/4
π/2
3π/4
π
θ
5π/4
3π/2
7π/4
2π

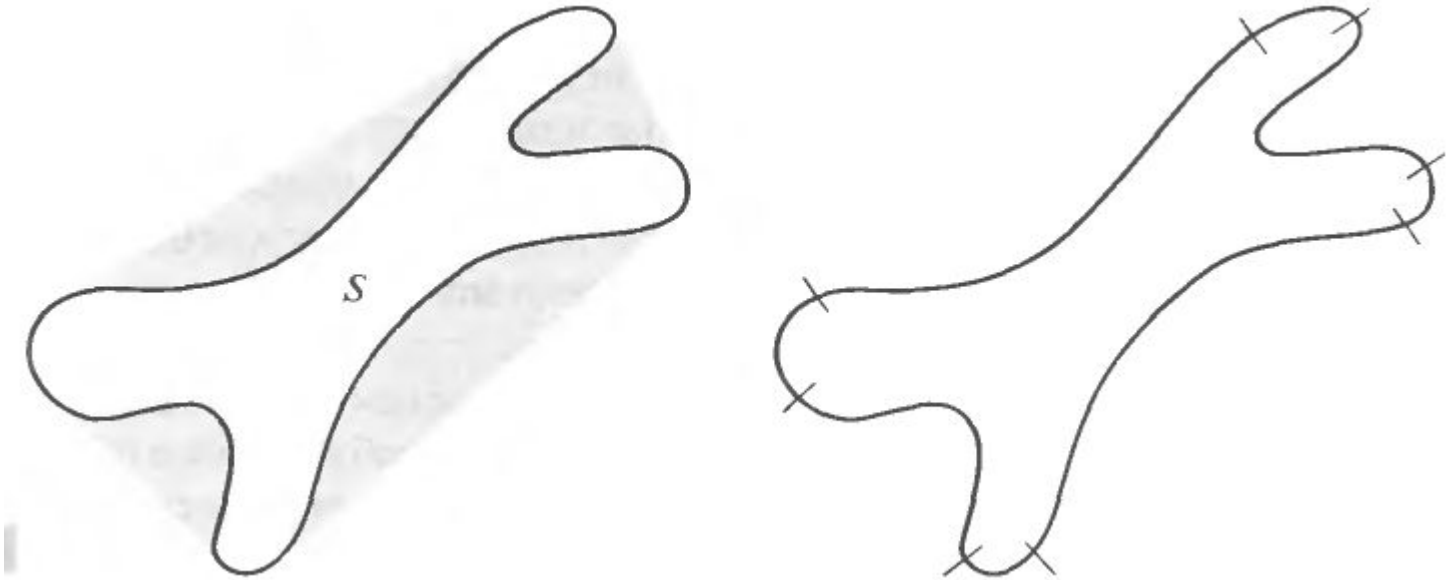
Сегменты границы

Хорошо применим, когда граница содержит одну или несколько хорошо выраженных вогнутостей, несущих информацию о форме объекта.

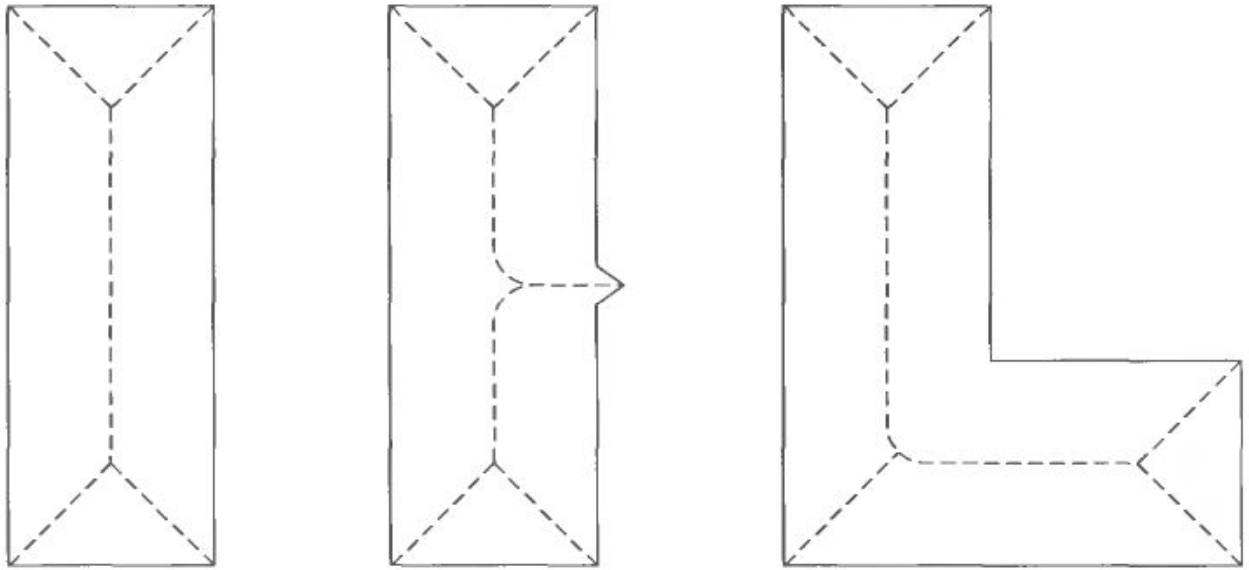
Сегменты границы

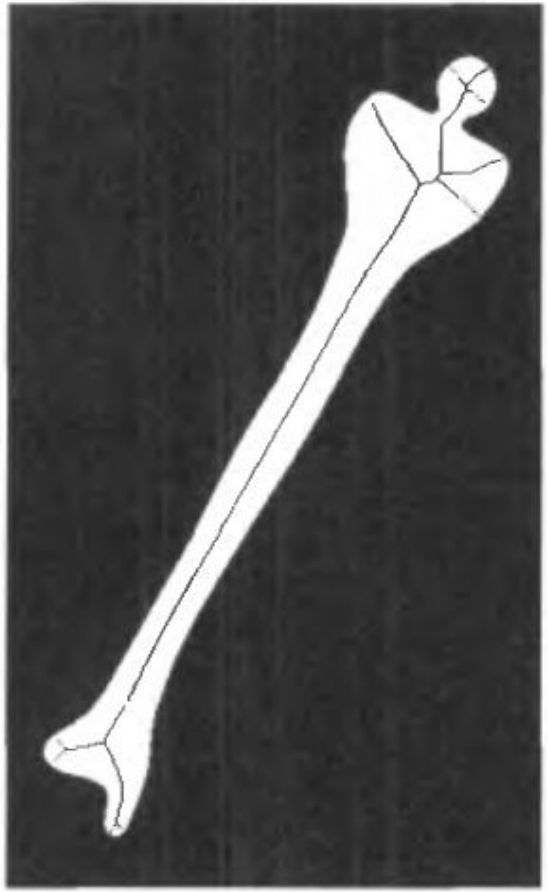
S - множество
H - выпуклая оболочка
H\S - дефект выпуклости *D* множества *S*

Сегменты границы



Остовы областей





Бедренная кость человека с наложенным остовом области.

Описание

Дескрипторы

Простые дескрипторы

- Длина
- Диаметр

$$Diam(B) = \max_{i,j}[D(p_i, p_j)]$$

- Эксцентриситет
- Кривизна