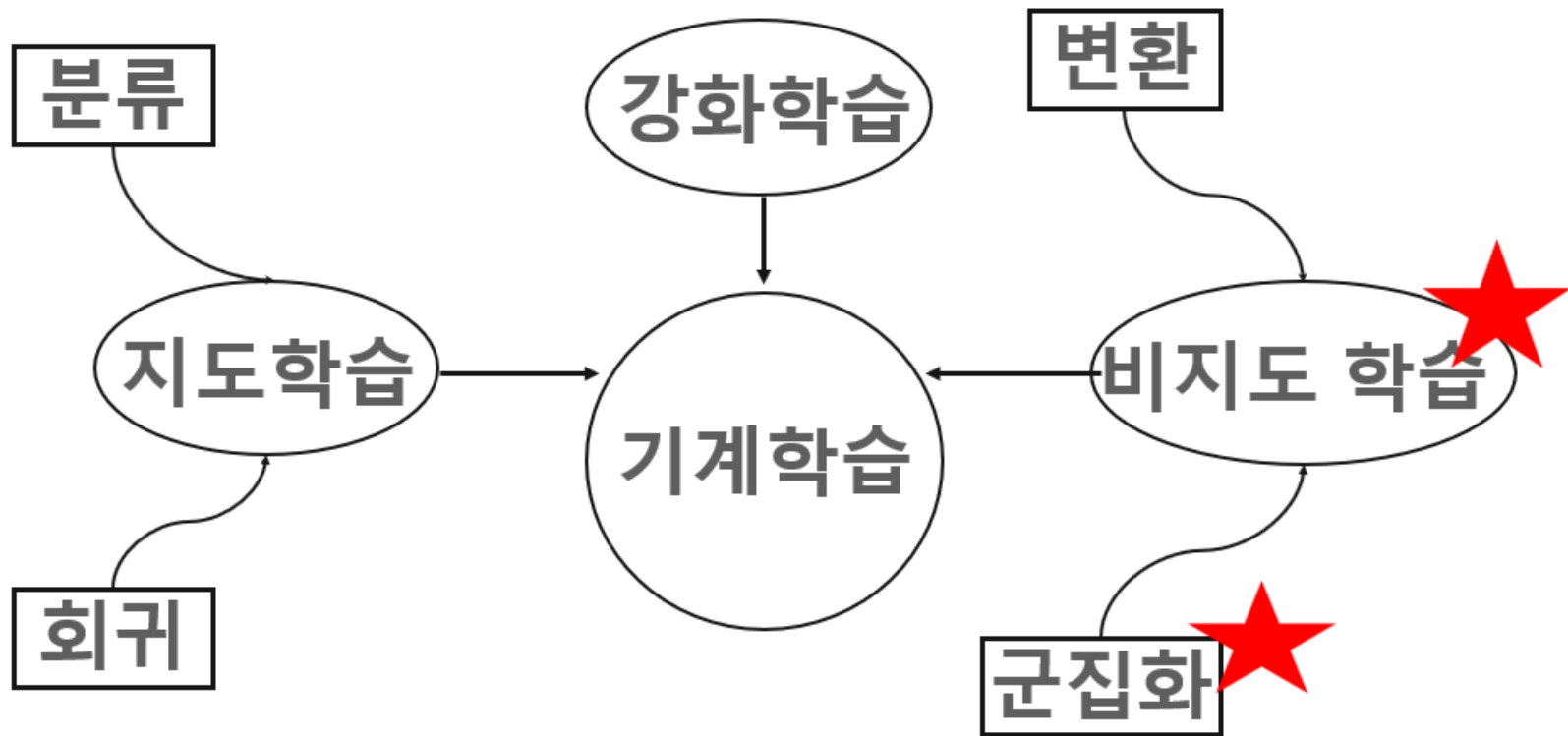


비지도학습(군집화)

홍익 대학교
Hyun-Sun Ryu

머신러닝의 종류



군 집화(Clustering)

군집화(Clustering)

- 주어진 데이터를 유사한 데이터들의 그룹으로 나누는 것을 군집화(clustering)이라 한다.

군집화(Clustering)



군집화(Clustering) vs. 분류

군집화
Clustering

분류
classification



군집화(Clustering)

군집화
Clustering

분류
classification

군집1
(cluster)



군집2
(cluster)



분류(Classification)

군집화
Clustering

분류
classification

개



고양이



참고: 군집화(clustering) vs. 분류(classification)

비지도

탐험

변수 | 변수 | 변수

지도학습

역사

독립변수 | 종속변수

참고: 군집화 vs. 연관규칙

군집화

clustering

연관규칙

association rule

군집화(Clustering)

- 군집화는 **많은 양의 관측치를 요약**
- 데이터를 조직화
- 또는 많은 데이터들 중에서 outlier를 판별
- 또, 많은 양의 데이터를 classification 하기 위한 전 단계의 작업
- 유의미하거나 유용한 데이터의 구조를 파악하여 활용하거나 데이터의 구조를 이해하는 목적으로 **분석 초기 탐색적 분석 단계**에서 아주 많이 활용

군집화(Clustering)의 활용

- 시장의 세그멘테이션 조사
- 문서 데이터 클러스터링
- 자율주행 자동차의 이미지 인식

세그멘테이션 (Segmentation)이란?

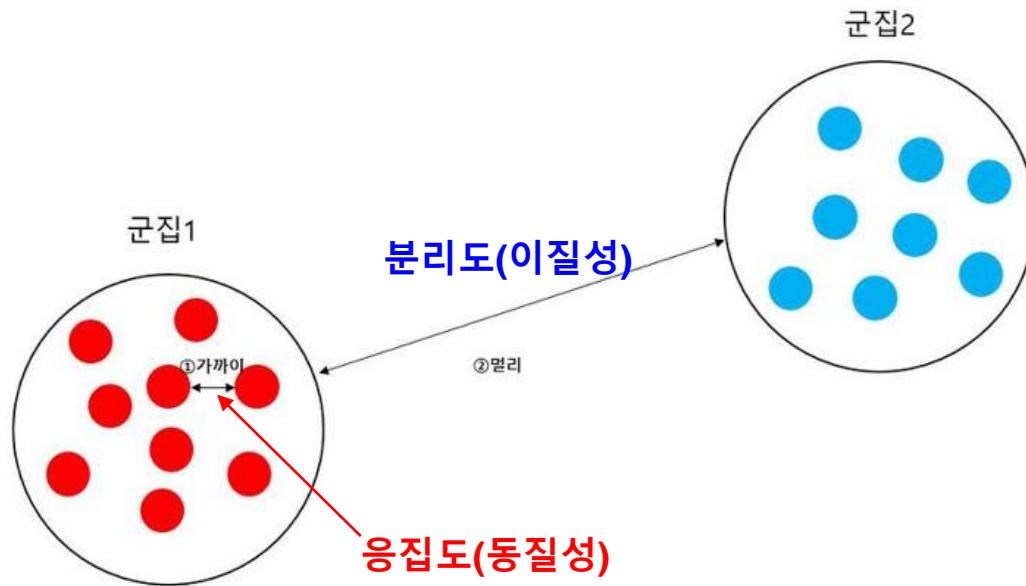
- 소비자를 비슷한 성향을 가진 사람들끼리 묶어 놓은 그룹
- 하나의 시장을 여러 고객의 하위군집으로 구분하는 것으로 구매패턴을 분석하여 비슷한 구매 성향을 가진 사람들을 군집으로 만듦으로서 마케팅 효과를 극대화

군집화(Clustering)의 평가척도

군집(cluster)내의 응집도(cohesion)를 최대화하고

군집간의 분리도(separation)를 최대화

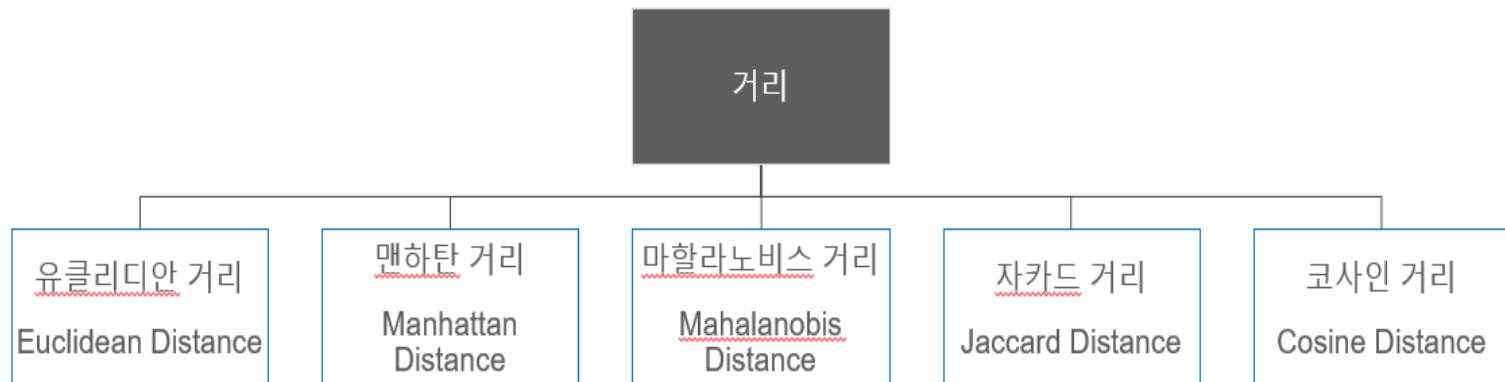
군집화(Clustering)



군집화의 비유사성(unsimilarity) 측도

거리

군집화의 비유사성 척도



군집화의 비유사성 측도

택시가 목적지까지 갈 때 건물과 장애물을 피해 길을 따라 가듯이
그림과 같이 놓여진 격자 무늬 도로 안에서 최단 루트를 찾는
방법

맨하탄거리
Manhattan Distance

유클리디안 거리
Euclidean Distance

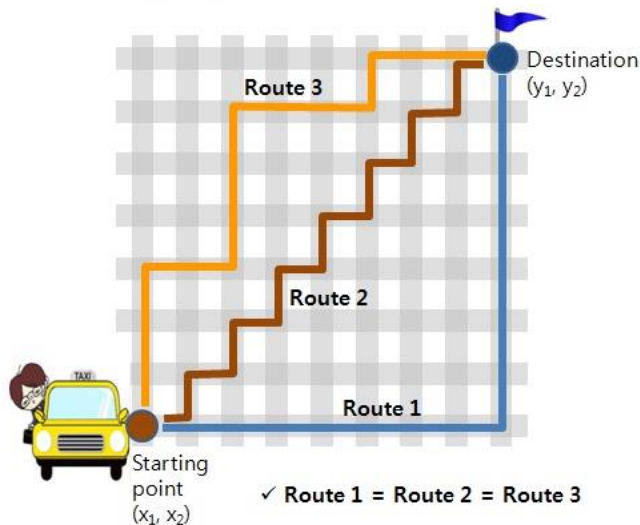
마할라노비스 거리
Mahalanobis Distance

자카드 거리
Jaccard Distance

코사인 거리
Cosine Distance

맨하탄 거리
(Manhattan Distance)

$$d_M(x, y) = \sum_{j=1}^m |x_j - y_j|$$



군집화의 비유사성 측도

유클리디안 거리는 두 점을 잇는 가장 짧은 직선 거리
(일반적으로 우리가 자를 대고 잴 거리)

맨하탄거리
Manhattan Distance

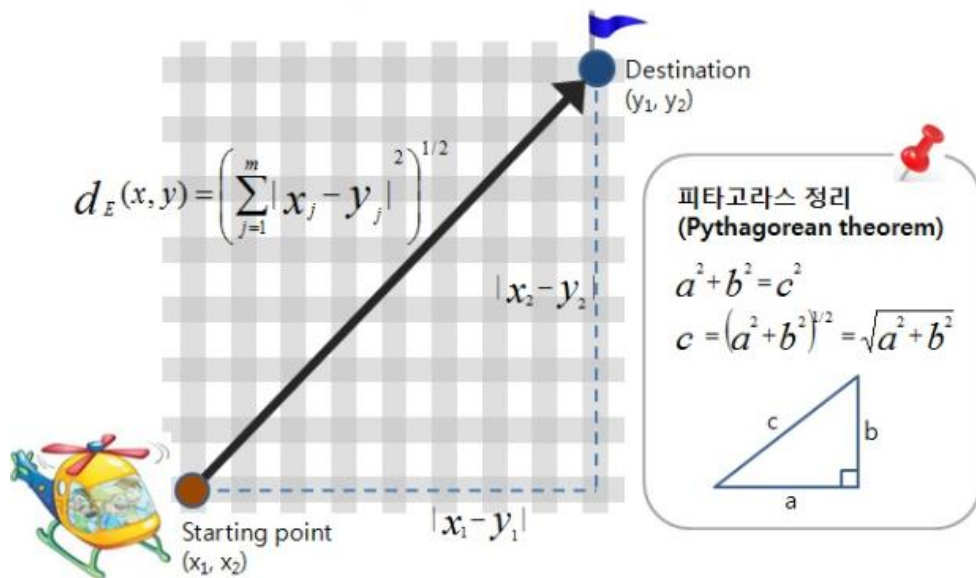
유클리디안 거리
Euclidean Distance

마할라노비스 거리
Mahalanobis Distance

자카드 거리
Jaccard Distance

코사인 거리
Cosine Distance

유클리드 거리 (Euclidean Distance)



군집화의 비유사성 측도

- 맨하탄 거리, 유클리디안 거리, 마할라노비스 거리는 **연속형 변수들 간의 거리**를 측정하는데 유용
- 마할라노비스 거리는 **평균과의 거리가 표준편차의 몇 배** 인지를 나타내는 값(표준정규분포로 바꾼 후 z값이 마할라노비스 거리)

맨하탄거리
Manhattan Distance

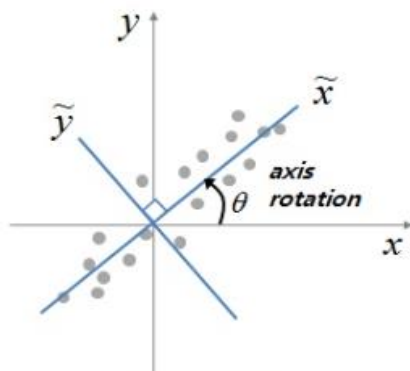
유클리디안 거리
Euclidean Distance

마할라노비스 거리
Mahalanobis Distance

자카드 거리
Jaccard Distance

코사인 거리
Cosine Distance

상관성을 고려해 회전시킨 축

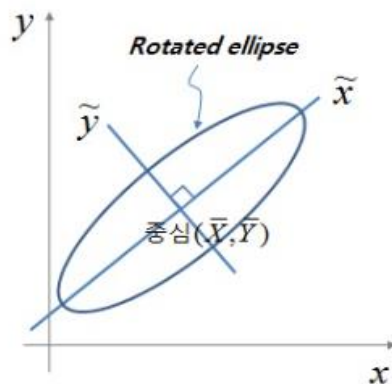


$$\begin{pmatrix} \tilde{x} \\ \tilde{y} \end{pmatrix} = \begin{pmatrix} \cos \theta & \sin \theta \\ -\sin \theta & \cos \theta \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} x \cos \theta + y \sin \theta \\ -x \sin \theta + y \cos \theta \end{pmatrix}$$

회전변환행렬

변수가 2개인 경우 두 개체간의 마할라노비스 거리

⇒ 상관성을 고려해 회전시킨 타원 그림



$$Z = \frac{X - \mu}{\sigma}$$

군집화의 비유사성 척도

- 자카드 거리 척도는 범주형 데이터에 대해서 비유사성을 측정하는 지표
- 자카드 거리는 전체 속성 중에 공통의 속성 비율이 얼마인지를 나타냄

맨하탄거리
Manhattan Distance

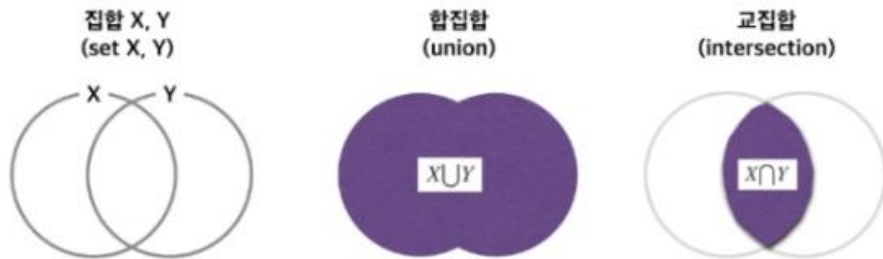
유클리디안 거리
Euclidean Distance

마할라노비스 거리
Mahalanobis Distance

자카드 거리
Jaccard Distance

코사인 거리
Cosine Distance

Jaccard Index & Jaccard Distance



유사성 측정

Jaccard index
(Intersection over Union,
Jaccard **similarity** coefficient,
Jaccard coefficient)

$$J(X, Y) = \frac{|X \cap Y|}{|X \cup Y|} = \frac{|X \cap Y|}{|X| + |Y| - |X \cap Y|}, \quad 0 \leq J(X, Y) \leq 1$$

비유사성 측정

Jaccard distance
(Jaccard **dissimilarity** coefficient)

$$d_{jaccard}(X, Y) = 1 - J(X, Y) = 1 - \frac{|X \cap Y|}{|X \cup Y|} = \frac{|X \cup Y| - |X \cap Y|}{|X \cup Y|}$$

군집화의 비유사성 척도

- 코사인 거리는 두 벡터 값에서 코사인 각도를 구하는 방법
- 문서를 유사도를 기준으로 분류 혹은 그룹핑 할 때 유용하게 사용

맨하탄거리
Manhattan Distance

유클리디안 거리
Euclidean Distance

마할라노비스 거리
Mahalanobis Distance

자카드 거리
Jaccard Distance

코사인 거리
Cosine Distance

코사인 유사도 (Cosine Similarity) vs. 코사인 거리 (Cosine Distance)

Cosine similarity

$$\text{cosine similarity} = \frac{X \cdot Y}{\|X\|_2 \cdot \|Y\|_2}$$


Cosine distance

$$d_{\text{cosine}}(X, Y) = 1 - \frac{X \cdot Y}{\|X\|_2 \cdot \|Y\|_2}, \text{ where } \|X\|_2 \text{ is the L2 norm}$$



문서 분류, 군집화 (Text Classification, Clustering) 에 활용

오렌지3에서 거리 확인하기

위젯	설명	입력	출력
 Distances	쌍별 거리 행렬을 계산한다.	Data	Distances

- Distances 위젯은 데이터 세트의 행 또는 열 사이의 거리를 계산
- 기본적으로 데이터는 개별 features를 동일하게 취급하도록 정규화
- 정규화는 항상 열별로 수행

오렌지3에서 거리 확인하기

△ Distances ? ×

Distances between ①

☒ Rows
☐ Columns

Distance Metric ②

Cosine ▾

☐ Normalized

☒ Apply Automatically

? 📄

① Distances between

행 사이의 거리를 측정할 것인지 또는 열 사이의 거리를 측정할 것인지 여부를 선택한다.

거리 메트릭을 선택한다.

Euclidean: "직선", 두 점 사이의 거리

Manhattan: 모든 속성에 대한 절대적 차이의 합계

Cosine: 내부 제품 공간의 두 벡터 간 각도의 코사인

Jaccard: 교차점 크기를 표본 집합의 결합 크기로 나눈 값

Spearman: 값 순위 간의 선형 상관 관계, [0, 1] 간격의 거리로 다시 매핑

Spearman absolute: 절대값 순위 간의 선형 상관 관계, [0, 1] 간격의 거리로 다시 매핑

Pearson: 값 사이의 선형 상관 관계, [0, 1] 간격의 거리로 다시 매핑됨

Pearson absolute: 절대값 사이의 선형 상관 관계, [0, 1] 간격의 거리로 다시 매핑


Hamming: 해당 값이 다른 features 수

Bhattacharyya distance: 삼각 부등식을 따르지 않기 때문에 실제 거리가 아니라 두 확률 분포 사이의 유사성

Normalized: feature를 정규화한다. 정규화는 항상 열별로 수행된다. 값이 0 중심화되어 있고 크기가 조정된다. 결측값이 있는 경우 위젯은 행 또는 열의 평균값을 자동으로 귀속시킨다. 위젯은 숫자 및 범주형 데이터 모두에 대해 작동한다. 범주형 데이터의 경우 두 값이 동일하면 거리는 0이고('녹색' 및 '녹색') 값이 동일하지 않으면 1이다('녹색' 및 '청색').

② Distance Metric

오렌지3에서 거리 확인하기

위젯	설명	입력	출력
 Distance Matrix	거리 행렬을 표시한다.	Distances	Distances, Selected Data

- 거리 행렬은 집합 요소 간에 쌍으로 이동한 거리를 포함하는 2차원 배열
- 데이터 세트의 요소 수는 행렬의 크기를 정의
- 데이터 행렬은 계층적 군집화에 필수적이며, 그것들은 좌표 독립적인 방식으로 단백질 구조를 나타내는 데 사용되는 생물 정보학에서도 매우 유용

오렌지3에서 거리 확인하기

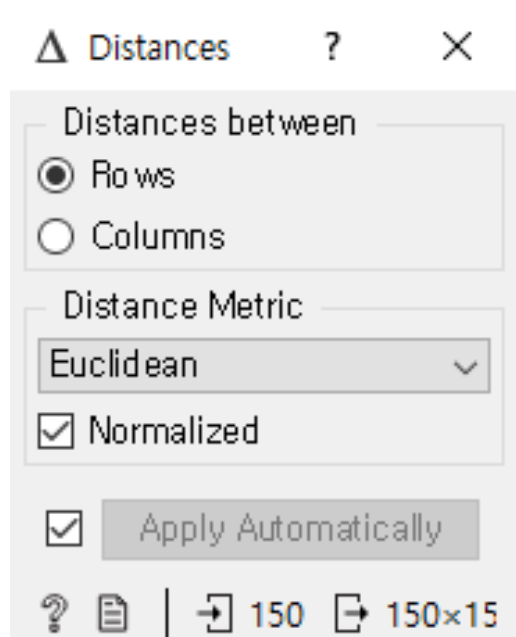
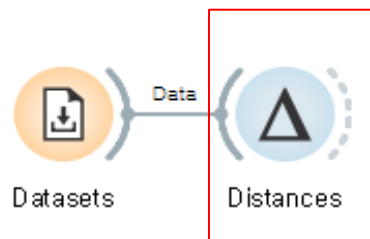
Distance Matrix

①	Iris-setosa	Iris-setosa	Iris-setosa	Iris-setosa	Iris-setosa	Iris-setosa	Iris-setosa	Iris-setosa	Iris-setosa	Iris-setosa
Iris-setosa	0.000	0.002	0.000	0.001	0.000	0.001	0.001	0.000	0.001	0.000
Iris-setosa	0.000	0.001	0.000	0.001	0.001	0.001	0.002	0.000	0.000	0.000
Iris-versicolor	0.072	0.060	0.070	0.058	0.075	0.064	0.069	0.065	0.057	0.060
Iris-versicolor	0.074	0.064	0.072	0.060	0.076	0.065	0.070	0.067	0.060	0.063
Iris-versicolor	0.082	0.070	0.081	0.067	0.085	0.074	0.080	0.075	0.067	0.070
Iris-versicolor	0.094	0.080	0.092	0.078	0.097	0.085	0.091	0.086	0.077	0.080
Iris-versicolor	0.087	0.074	0.085	0.071	0.090	0.078	0.084	0.079	0.071	0.073
Iris-versicolor	0.093	0.081	0.091	0.076	0.095	0.082	0.088	0.084	0.077	0.080
Iris-versicolor	0.081	0.071	0.079	0.066	0.083	0.070	0.076	0.073	0.066	0.070
Iris-versicolor	0.067	0.057	0.066	0.053	0.070	0.059	0.064	0.060	0.053	0.056
Iris-versicolor	0.080	0.067	0.078	0.065	0.083	0.072	0.078	0.072	0.065	0.068
Iris-versicolor	0.084	0.074	0.082	0.069	0.086	0.073	0.079	0.076	0.069	0.073
Iris-versicolor	0.088	0.074	0.087	0.073	0.092	0.081	0.087	0.081	0.072	0.074
Iris-versicolor	0.076	0.066	0.075	0.062	0.079	0.067	0.072	0.069	0.062	0.066
Iris-versicolor	0.086	0.071	0.084	0.071	0.091	0.080	0.086	0.079	0.070	0.077
Iris-versicolor	0.091	0.080	0.090	0.075	0.094	0.081	0.087	0.083	0.075	0.079
Iris-versicolor	0.060	0.051	0.058	0.047	0.062	0.051	0.056	0.054	0.047	0.050

Labels: iris ② ☒ Send Automatically

① Elements	데이터 세트의 요소와 요소 사이의 거리이다.
② Label	테이블에 레이블을 붙인다. 옵션은 변수에 따라 none 또는 열거형으로 제시된다.

오렌지3에서 거리 확인하기



오렌지3에서 거리 확인하기



Distance Matrix

	Iris-setosa	Iris-setosa	Iris-setosa	Iris-setosa	Iris-setosa	Iris-setosa	Iris-setosa	Iris-setosa	Iris-setosa	Iris-setosa	Iris-setosa
Iris-setosa		0.836	0.600	0.783	0.185	0.737	0.468	0.189	1.150	0.684	0.418
Iris-setosa	0.836		0.372	0.307	0.985	1.550	0.709	0.661	0.459	0.192	0.315
Iris-setosa	0.600	0.372		0.201	0.704	1.316	0.353	0.424	0.556	0.267	0.863
Iris-setosa	0.783	0.307	0.201		0.888	1.491	0.501	0.599	0.372	0.273	0.258
Iris-setosa	0.185	0.985	0.704	0.888		0.638	0.483	0.330	1.255	0.829	1.001
Iris-setosa	0.737	1.550	1.316	1.491	0.638		1.078	0.910	1.860	1.408	1.589
Iris-setosa	0.468	0.709	0.353	0.501	0.483	1.078		0.357	0.841	0.586	1.787
Iris-setosa	0.189	0.661	0.424	0.599	0.330	0.910	0.357		0.967	0.507	0.702
Iris-setosa	1.150	0.459	0.556	0.372	1.255	1.860	0.841	0.967		0.549	0.684
Iris-setosa	0.684	0.192	0.267	0.273	0.829	1.408	0.586	0.507	0.549		1.203
Iris-setosa	0.418	1.224	1.018	1.197	0.382	0.385	0.849	0.599	1.565	1.075	
Iris-setosa	0.315	0.665	0.359	0.521	0.378	0.985	0.211	0.176	0.891	0.508	0.891
Iris-setosa	0.863	0.126	0.353	0.258	1.001	1.589	0.702	0.684	0.391	0.189	
Iris-setosa	1.078	0.536	0.490	0.357	1.160	1.787	0.737	0.907	0.239	0.563	0.239
Iris-setosa	1.018	1.811	1.613	1.800	0.951	0.468	1.427	1.203	2.164	1.669	1.669
Iris-setosa	1.571	2.399	2.152	2.334	1.452	0.861	1.891	1.753	2.702	2.252	2.252

Labels: Iris

☒ Send Automatically

오렌지3에서 거리 확인하기

Distance Matrix

	Iris-setosa	Iris-setosa	Iris-setosa	Iris-setosa	Iris-setosa	Iris-setosa	Iris-setosa	Iris-setosa	Iris-setosa	Iris-setosa	Iris-setosa
Iris-setosa		0.836	0.600	0.783	0.185	0.737	0.468	0.189	1.150	0.684	0.418
Iris-setosa	0.836		0.372	0.307	0.985	1.550	0.709	0.661	0.459	0.192	0.122
Iris-setosa	0.600	0.372		0.201	0.704	1.316	0.353	0.424	0.556	0.267	1.018
Iris-setosa	0.783	0.307	0.201		0.888	1.491	0.501	0.599	0.372	0.273	0.357
Iris-setosa	0.185	0.985	0.704	0.888		0.638	0.483	0.330	1.255	0.829	0.382
Iris-setosa	0.737	1.550	1.316	1.491	0.638		1.078	0.910	1.860	1.408	0.385
Iris-setosa	0.468	0.709	0.353	0.501	0.483	1.078		0.357	0.841	0.586	0.849
Iris-setosa	0.189	0.661	0.424	0.599	0.330	0.910	0.357		0.967	0.507	0.599
Iris-setosa	1.150	0.459	0.556	0.372	1.255	1.860	0.841	0.967		0.549	1.565
Iris-setosa	0.684	0.192	0.267	0.273	0.829	1.408	0.586	0.507	0.549		1.075
Iris-setosa	0.418	1.224	1.018	1.197	0.382	0.385	0.849	0.599	1.565	1.075	
Iris-setosa	0.315	0.665	0.359	0.521	0.378	0.985	0.211	0.176	0.891	0.508	0.891
Iris-setosa	0.863	0.126	0.353	0.258	1.001	1.589	0.702	0.684	0.391	0.189	0.391
Iris-setosa	1.078	0.536	0.490	0.357	1.160	1.787	0.737	0.907	0.239	0.563	0.907
Iris-setosa	1.018	1.811	1.613	1.800	0.951	0.468	1.427	1.203	2.164	1.669	1.203
Iris-setosa	1.571	2.399	2.152	2.334	1.452	0.861	1.891	1.753	2.702	2.252	1.753

Labels: Iris ☒ Send Automatically

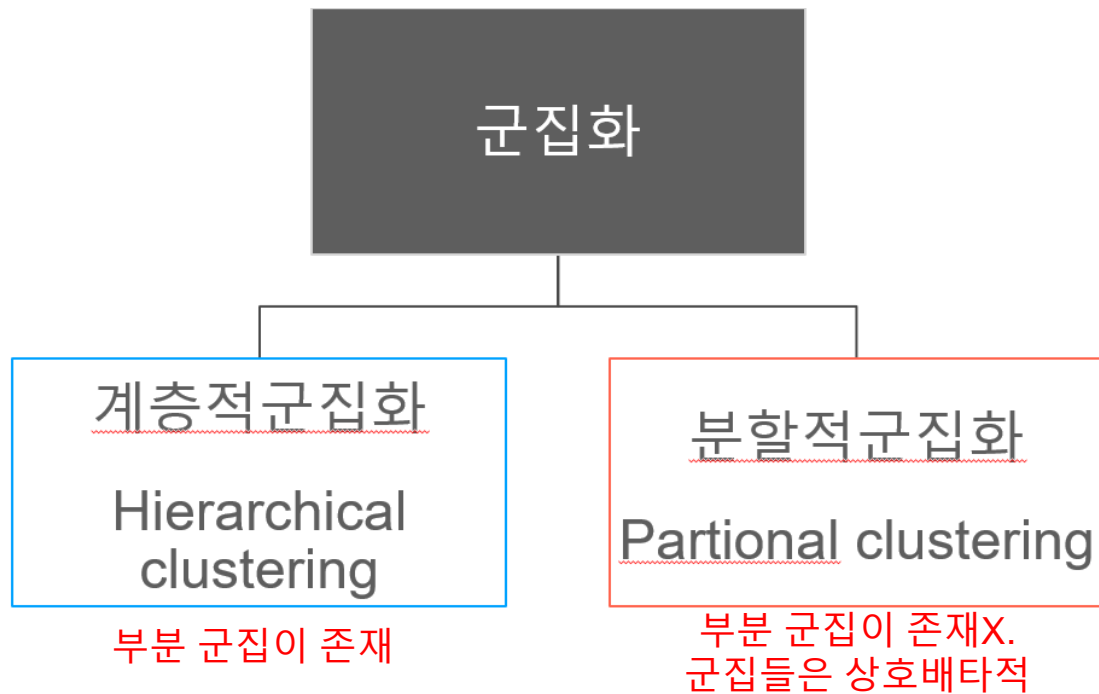
Distance Matrix

	Iris-setosa	Iris-setosa	Iris-setosa	Iris-setosa	Iris-setosa	Iris-setosa	Iris-setosa	Iris-setosa	Iris-setosa	Iris-setosa	Iris-setosa
Iris-setosa		0.001	0.000	0.001	0.000	0.001	0.001	0.000	0.001	0.001	0.001
Iris-setosa	0.001		0.001	0.001	0.003	0.003	0.004	0.001	0.001	0.000	0.000
Iris-setosa	0.000	0.001		0.001	0.000	0.001	0.001	0.000	0.001	0.001	0.001
Iris-setosa	0.001	0.001	0.001		0.001	0.001	0.001	0.000	0.000	0.001	0.001
Iris-setosa	0.000	0.003	0.000	0.001		0.001	0.000	0.001	0.001	0.001	0.002
Iris-setosa	0.001	0.003	0.001	0.001	0.001		0.000	0.001	0.001	0.001	0.003
Iris-setosa	0.001	0.004	0.001	0.001	0.000	0.000		0.001	0.002	0.003	0.003
Iris-setosa	0.000	0.001	0.000	0.000	0.001	0.001	0.001		0.000	0.001	0.001
Iris-setosa	0.001	0.001	0.001	0.000	0.001	0.001	0.002	0.000		0.000	0.000
Iris-setosa	0.001	0.000	0.001	0.001	0.002	0.003	0.003	0.001	0.000		0.001
Iris-setosa	0.000	0.001	0.000	0.001	0.000	0.001	0.001	0.000	0.001	0.001	0.001
Iris-setosa	0.001	0.002	0.001	0.000	0.001	0.000	0.001	0.000	0.001	0.001	0.001
Iris-setosa	0.001	0.000	0.001	0.001	0.002	0.003	0.003	0.001	0.001	0.000	0.000
Iris-setosa	0.000	0.002	0.000	0.002	0.000	0.002	0.001	0.001	0.002	0.002	0.002
Iris-setosa	0.001	0.004	0.002	0.005	0.002	0.004	0.003	0.003	0.004	0.004	0.004
Iris-setosa	0.002	0.006	0.002	0.004	0.001	0.001	0.001	0.003	0.004	0.004	0.006

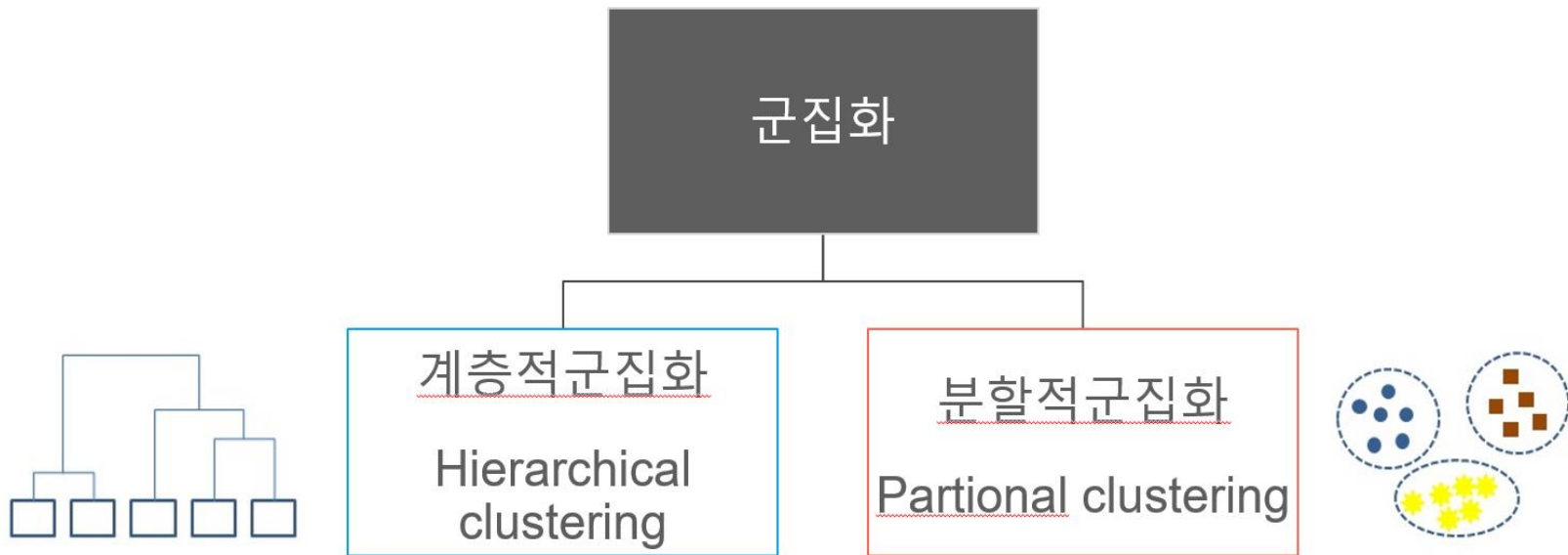
Labels: Iris ☒ Send Automatically

군집화(Clustering) 알고리즘

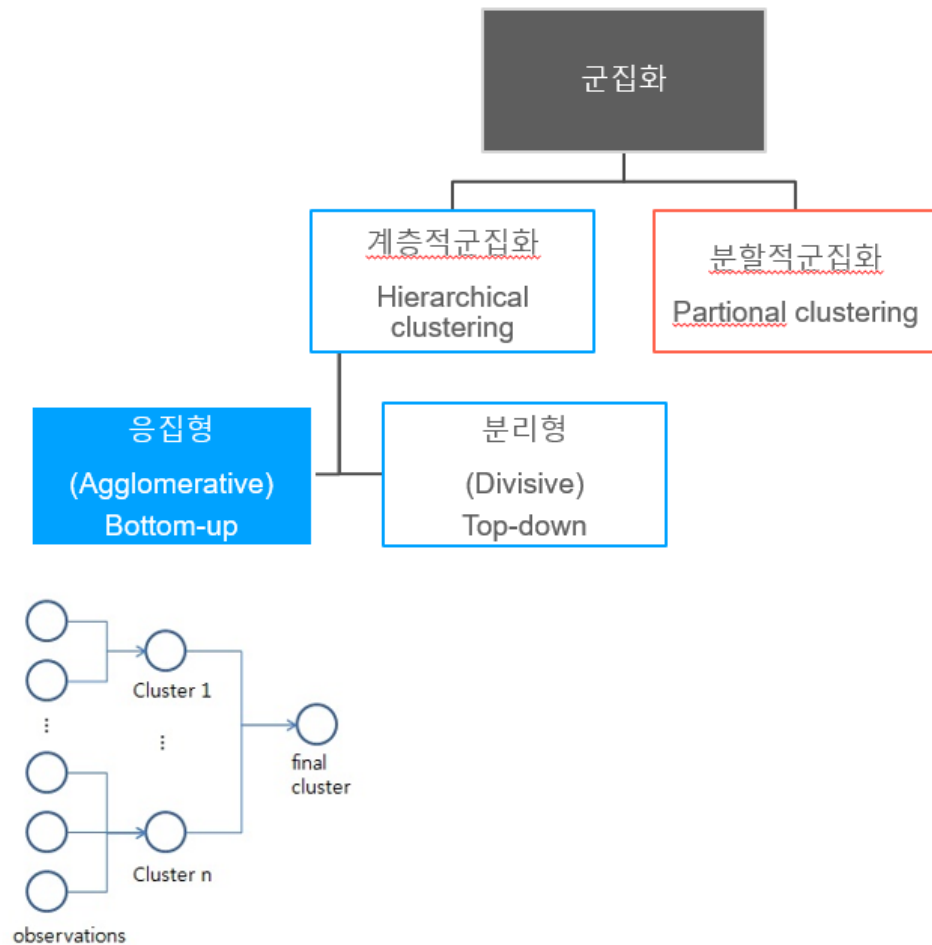
군집화 모델



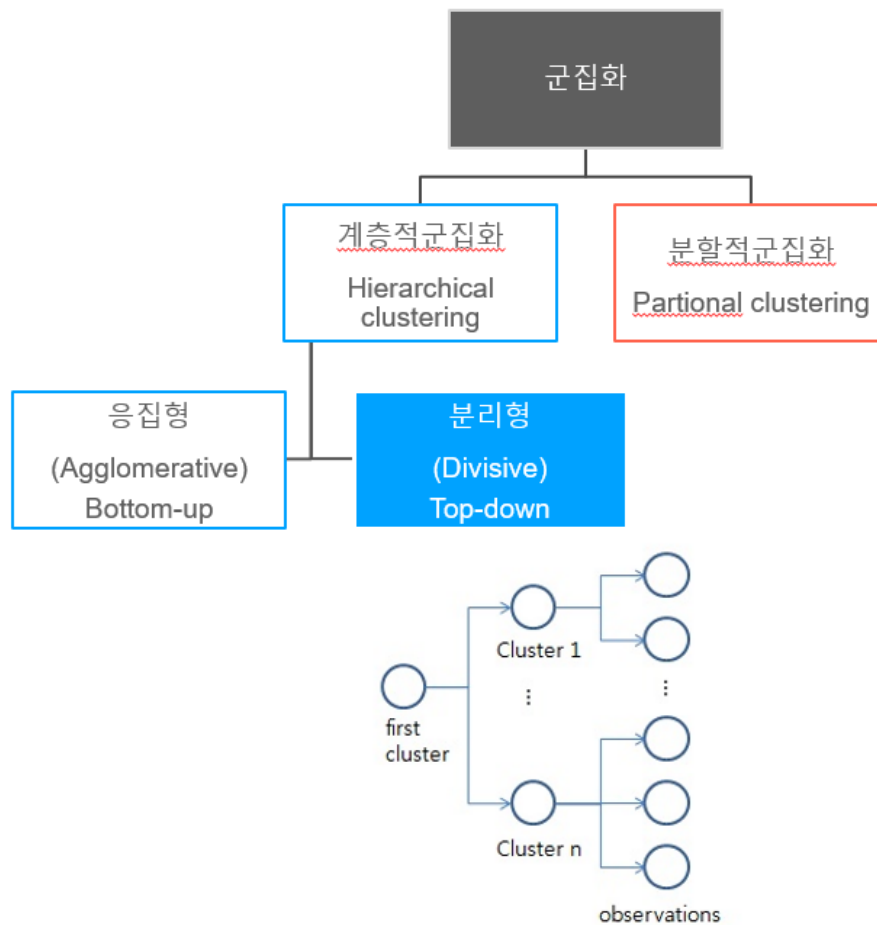
군집화 모델



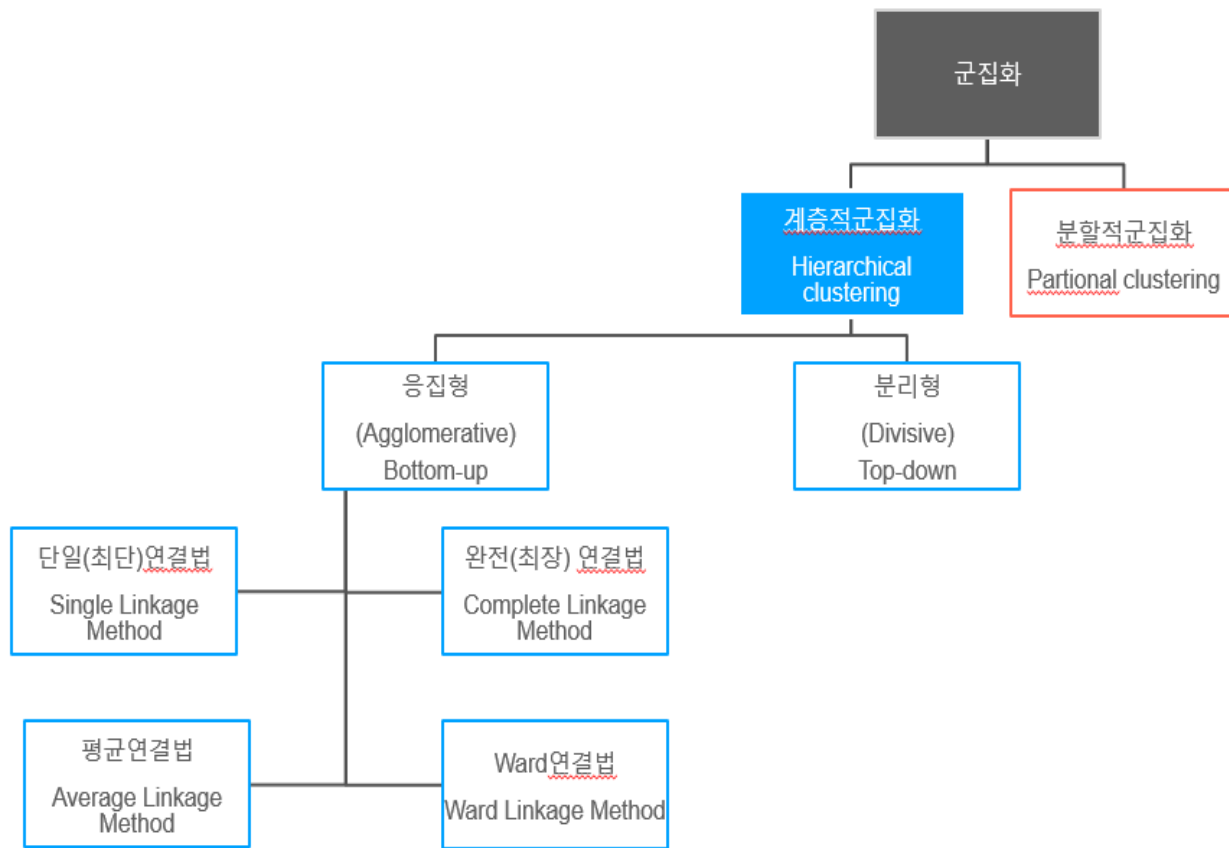
군집화 모델



군집화 모델

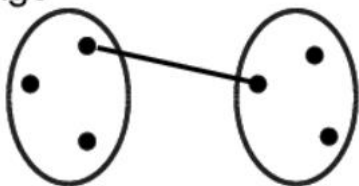


군집화 모델

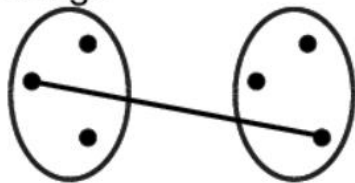


군집화 모델

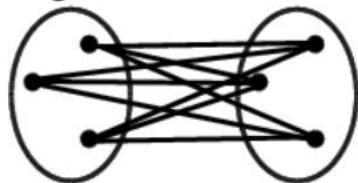
Single Linkage



Complete Linkage



Average Linkage



- **단일(최단)연결법:** 두 군집에 속한 **가장 가까운 두 점 사이의 거리**를 군집 간의 거리로 측정하는 방법
- **완전(최장)연결법:** 두 군집에 속한 **가장 가장 먼 두 점 사이의 거리**를 군집 간의 거리로 측정하는 방법
- **평균연결법:** 군집 안의 모든 데이터와 다른 군집 내 **모든 데이터 사이의 평균 거리**를 측정하는 방법으로 노이즈나 아웃라이너에 덜 취약
- **Ward연결법:** 두 군집을 합쳐졌을 때 증가하는 오차제곱합(Error Sum of Squares)의 증가분에 기반해 거리를 측정하는 방법(**분산을 가장 작게 증가시키는 두 군집을 합치는 방법**)

군집화 모델

Hierarchical
Clustering

K-means

DBSCAN

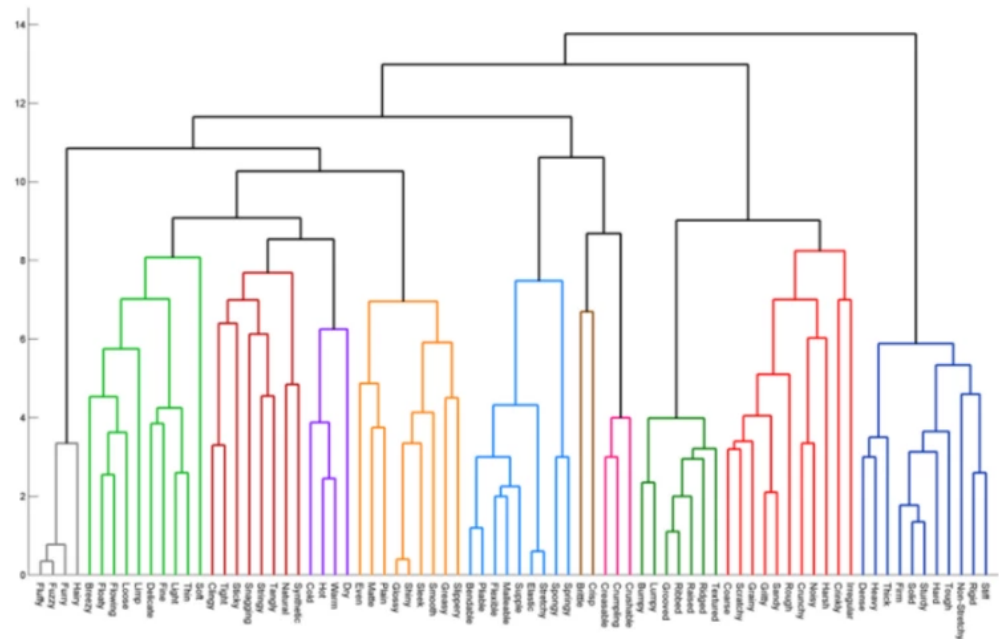
계층적 트리 모델을 이용하여 개별 개체들을
순차적, 계층적으로 유사한 개체 내지 그룹과
통합하여 군집화를 수행하는 알고리즘

군집화 모델

Hierarchical Clustering

K-means

DBSCAN




덴드로그램

군집화 모델

Hierarchical
Clustering

K-means

DBSCAN

위젯	설명	입력	출력
 Hierarchical Clustering	입력된 거리 행렬에서 생성된 계층적 군집화의 <u>덴드로그램</u> 을 표시한다.	Distances	Selected Data, Data

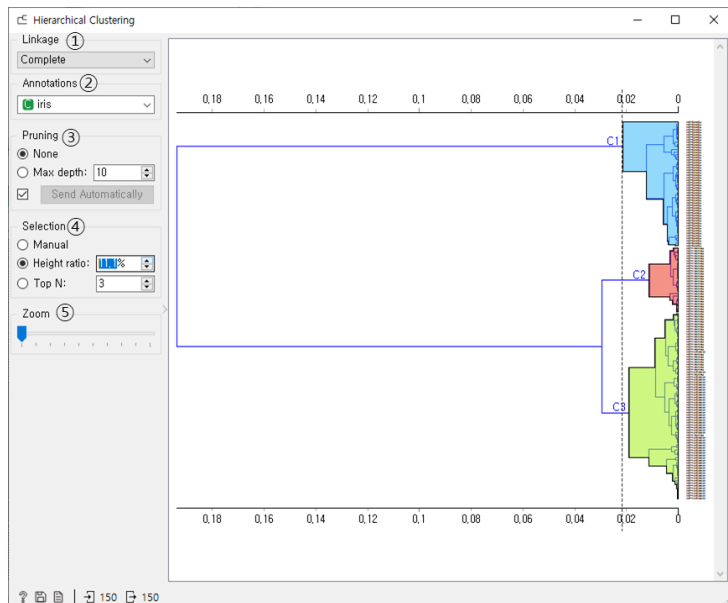
- Hierarchical Clustering 위젯은 거리 행렬(Distance Matrix)에서 임의의 개체 유형의 계층적 클러스터링을 계산하고 해당 덴드로그램을 표시

군집화 모델

Hierarchical Clustering

K-means

DBSCAN



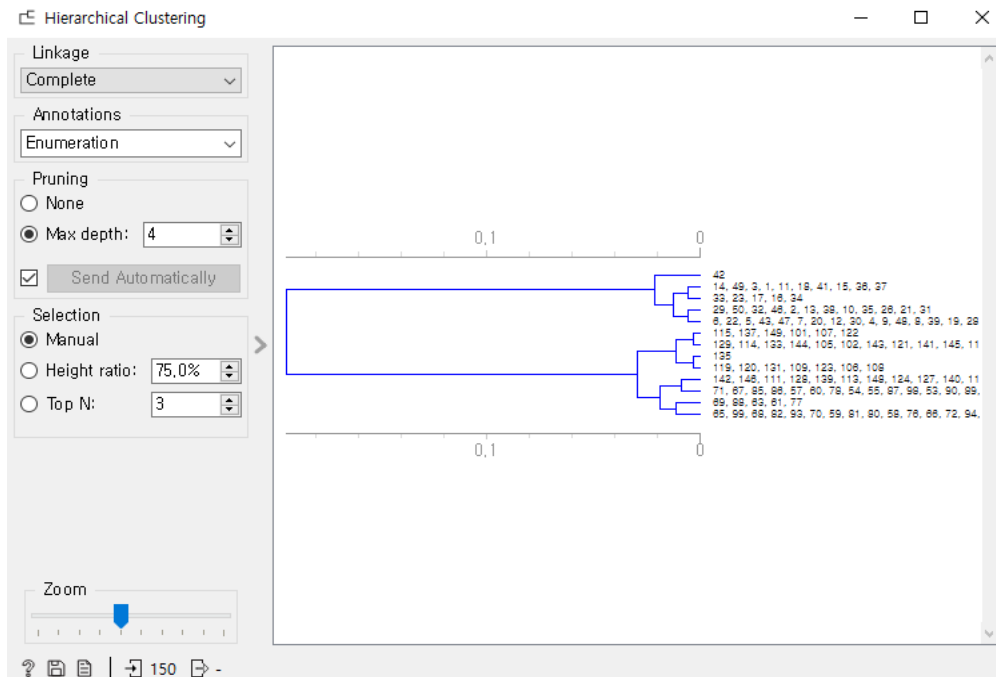
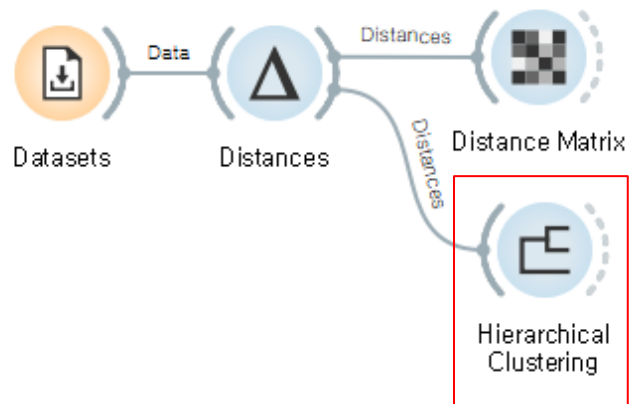
	<p>클러스터 간 거리를 측정하는 네 가지 방법을 지원한다.</p> <p>Single: 두 군집의 가장 가까운 요소 사이의 거리를 계산한다.</p> <p>Average: 두 군집의 요소 사이의 평균 거리를 계산한다.</p> <p>Weighted: WPGMA 방법을 사용한다.</p> <p>Complete: 군집의 가장 먼 요소 사이의 거리를 계산한다.</p> <p>Ward: 두 군집 간의 유사성을 두 군집이 합쳐졌을 때의 오차 제곱합(ESS)의 증가분에 기반해서 계산한다.</p>
① Linkage	
② Annotation	<p>덴드로그램의 노드 레이블은 Annotation 박스에서 선택할 수 있다.</p>
③ Pruning	<p>덴드로그램의 최대 깊이를 선택하여 가지치기 상자에서 큰 덴드로그램을 제거할 수 있다. 이는 실제 클러스터링이 아니라 디스플레이에만 영향을 미친다.</p>
④ Selection	<p>위젯은 세 가지 선택 방법을 제공한다.</p> <p>Manual: 덴드로그램 내부를 클릭하면 클러스터가 선택된다. Ctrl/Cmd를 눌러 여러 클러스터를 선택할 수 있다. 선택한 각 클러스터는 다른 색으로 표시되고 출력에서 별도의 클러스터로 처리된다.</p> <p>Height ratio: 덴드로그램의 아래쪽 또는 위쪽 눈금자를 클릭하면 그래프에 절단선이 배치된다. 줄의 오른쪽에 있는 항목이 선택된다.</p> <p>Top N: Top 노드 수를 선택한다.</p>
⑤ Zoom	<p>축소 및 스크롤을 사용하여 확대 또는 축소한다.</p>

군집화 모델

Hierarchical
Clustering

K-means

DBSCAN

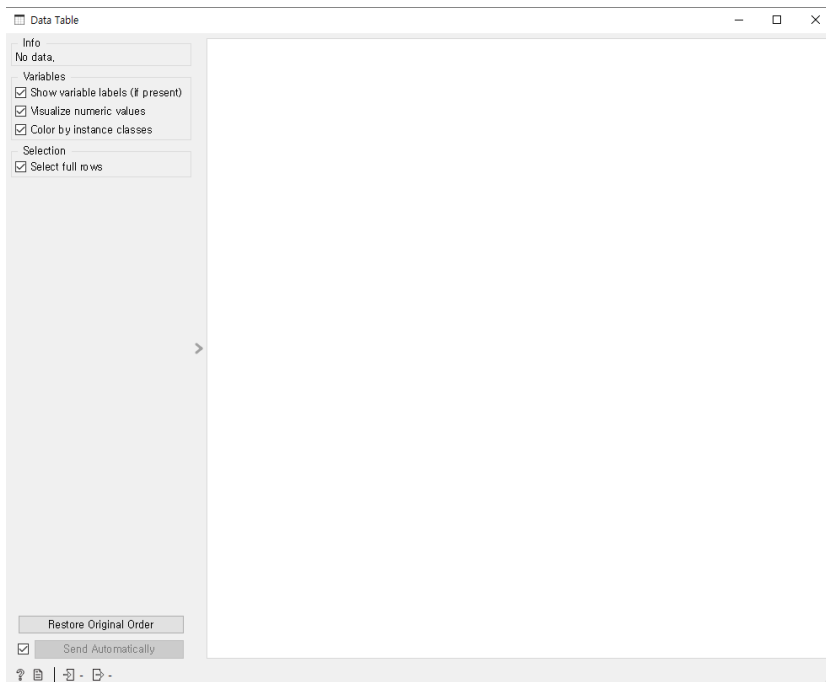
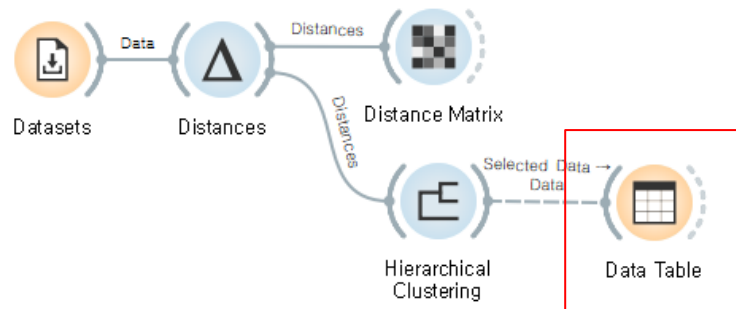


군집화 모델

Hierarchical
Clustering

K-means

DBSCAN

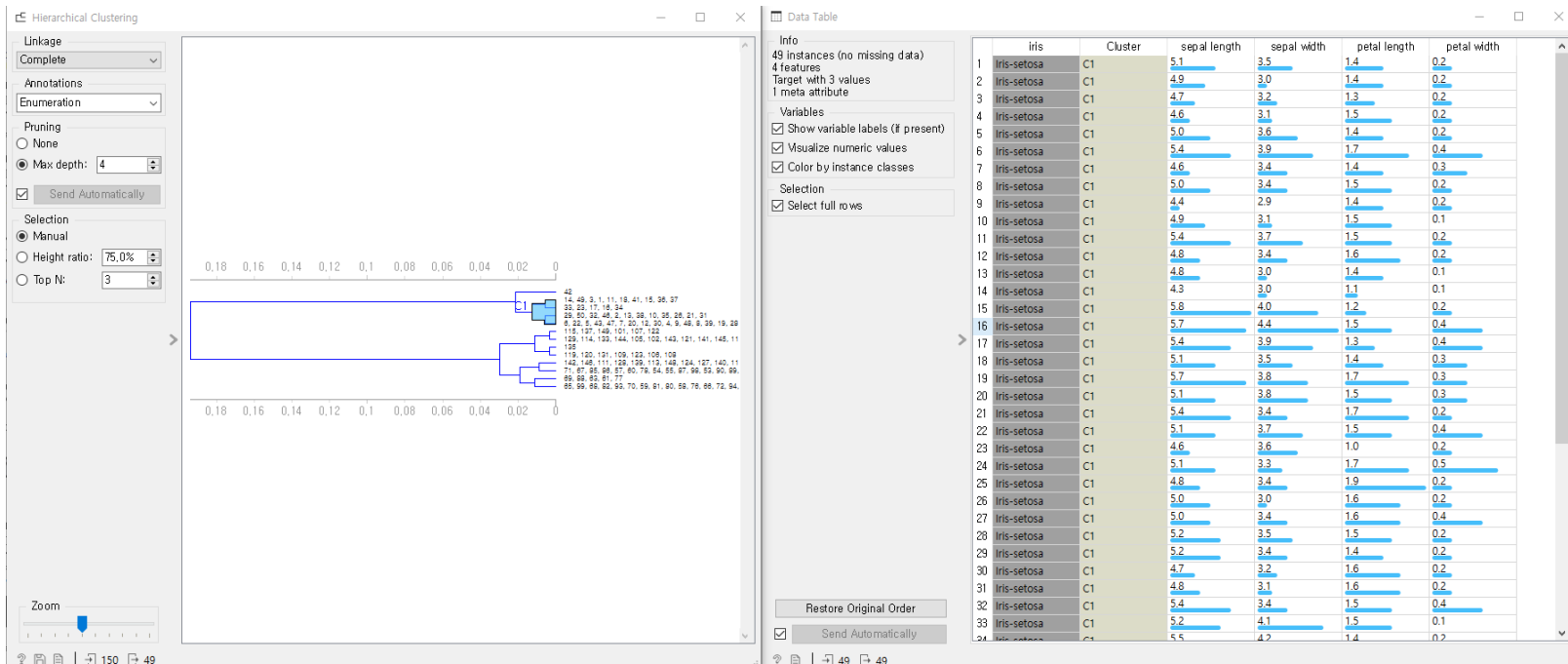


군집화 모델

Hierarchical Clustering

K-means

DBSCAN

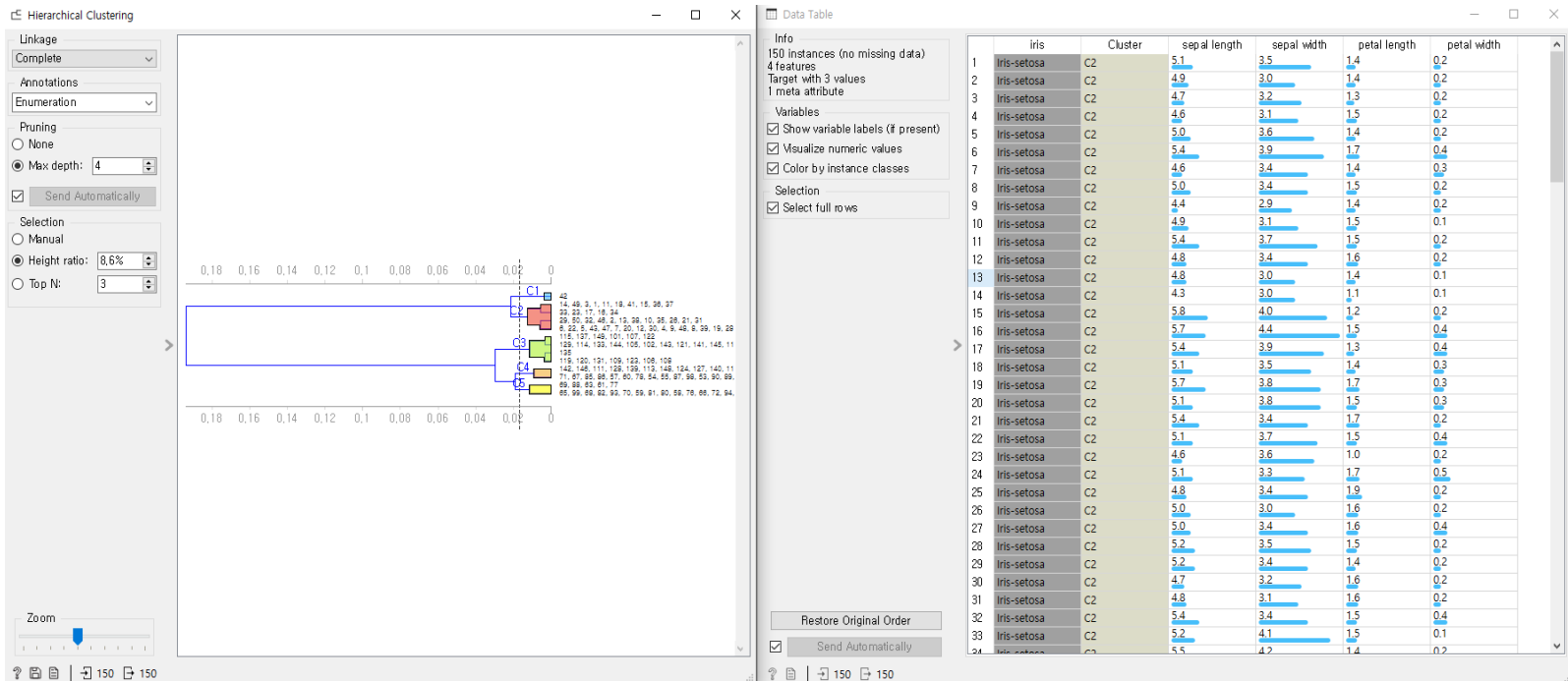


군집화 모델

Hierarchical Clustering

K-means

DBSCAN

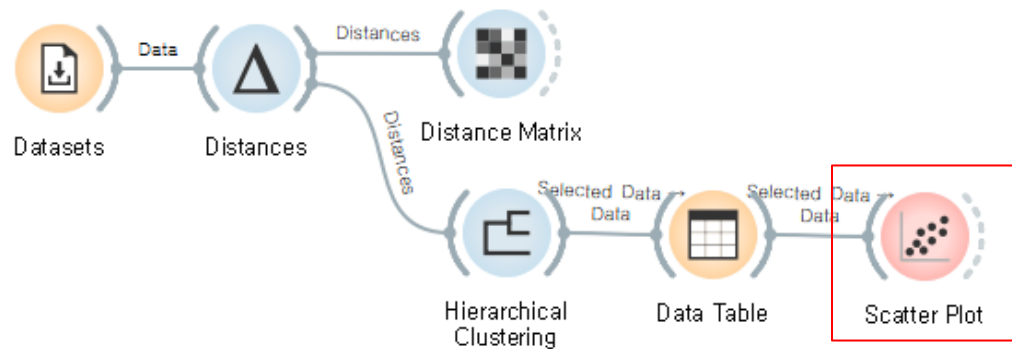


군집화 모델

Hierarchical
Clustering

K-means

DBSCAN

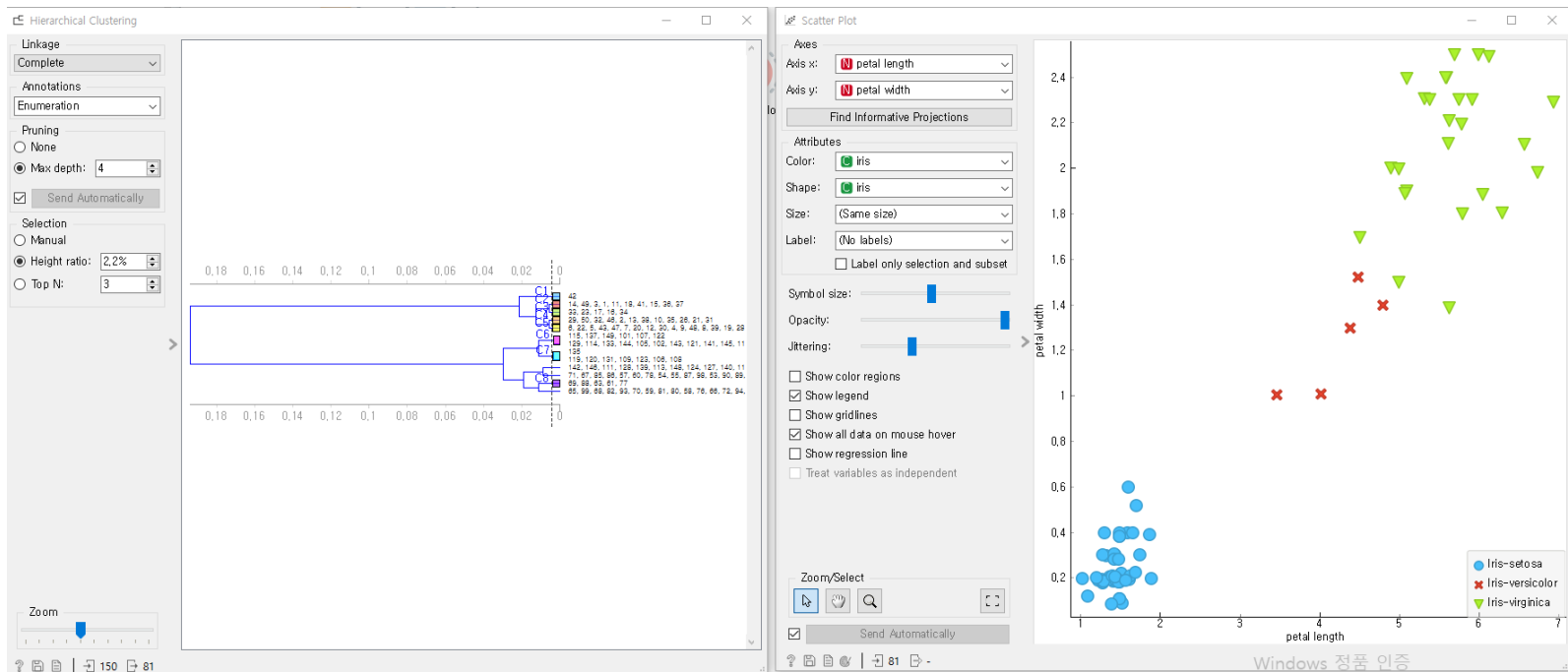


군집화 모델

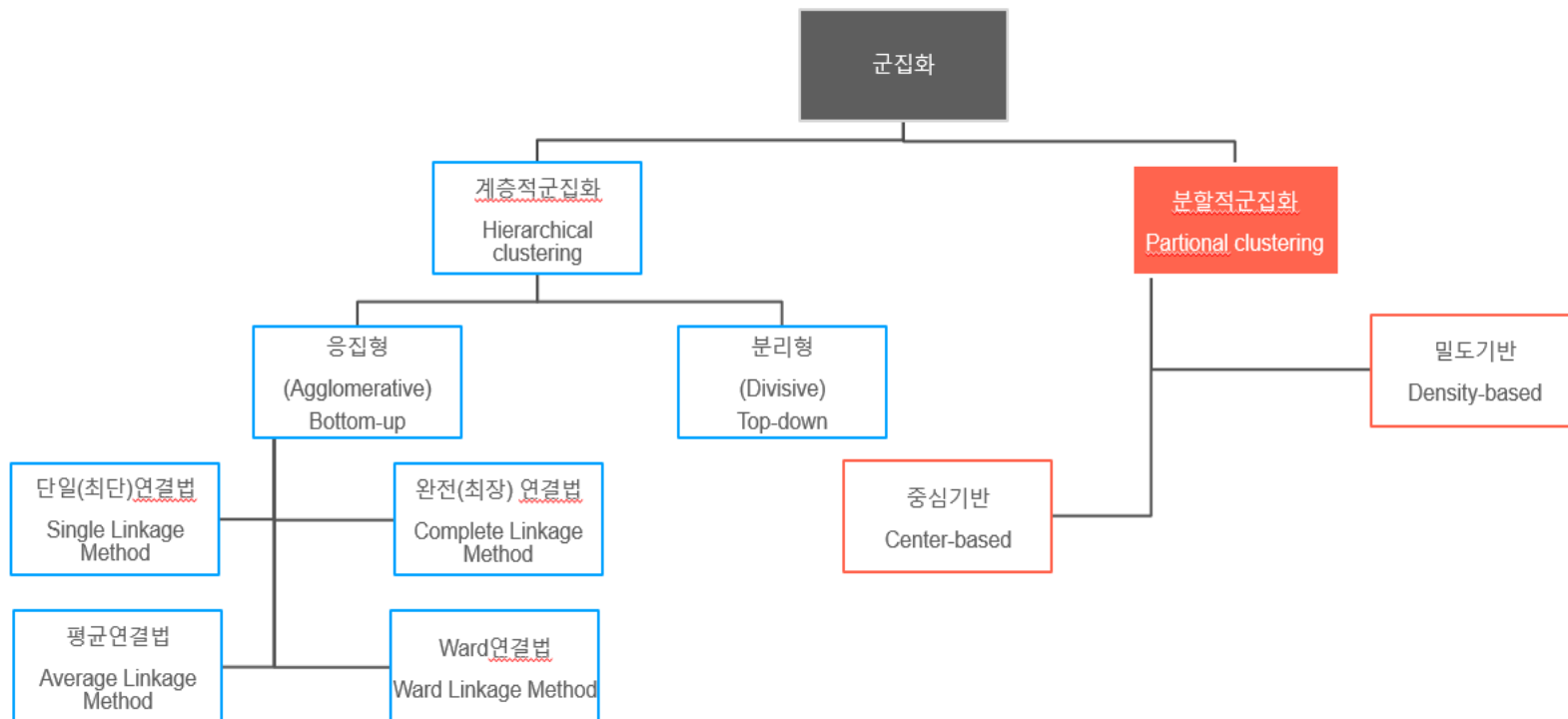
Hierarchical
Clustering

K-means

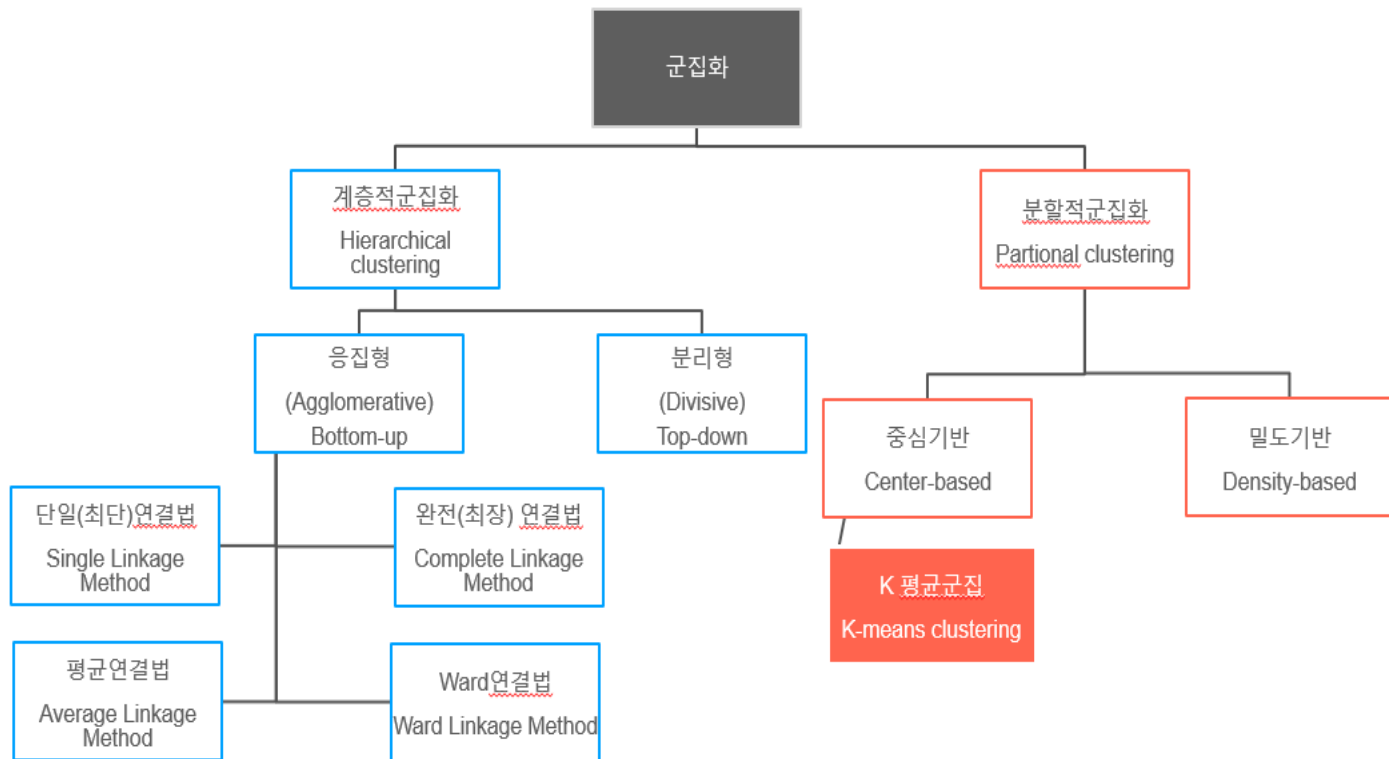
DBSCAN



군집화 모델



군집화 모델



K-Means

Hierarchical
Clustering

K평균 군집화
K-means

DBSCAN

군집 중심점(centroid)을 지정해 해당 중심에서 가장 가까운 점들을
선택하는 군집화 기법

K-Means

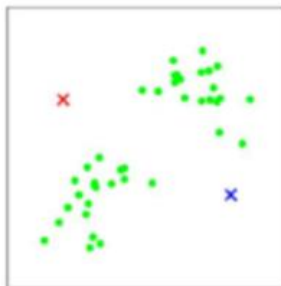
Hierarchical
Clustering

K평균 군집화
K-means

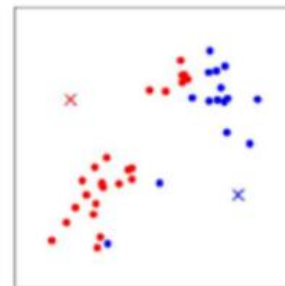
DBSCAN



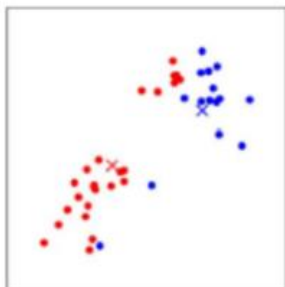
(a)



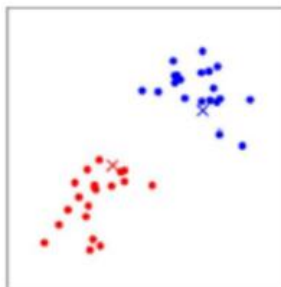
(b)



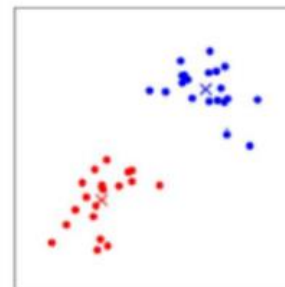
(c)



(d)



(e)



(f)

K-Means

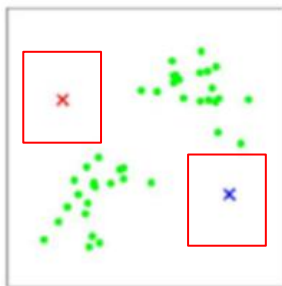
Hierarchical
Clustering

K평균 군집화
K-means

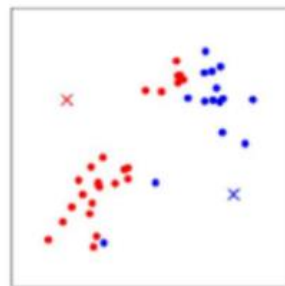
DBSCAN



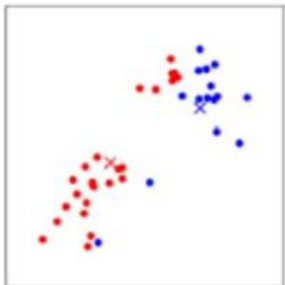
(a)



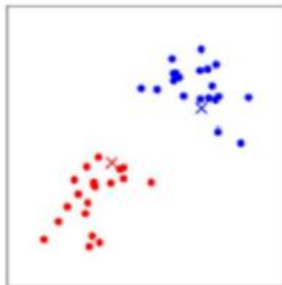
(b)



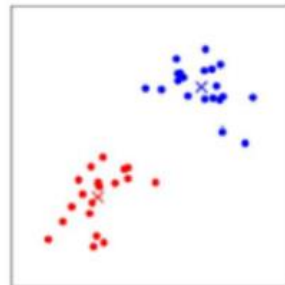
(c)



(d)



(e)



(f)

K-Means

Hierarchical
Clustering

K평균 군집화
K-means

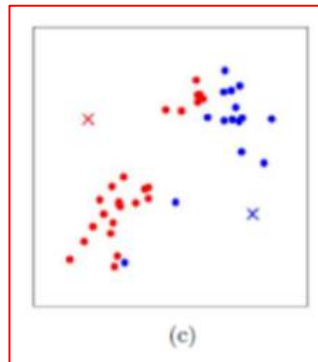
DBSCAN



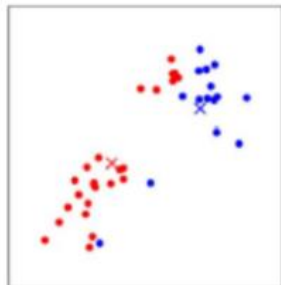
(a)



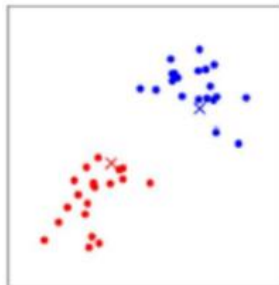
(b)



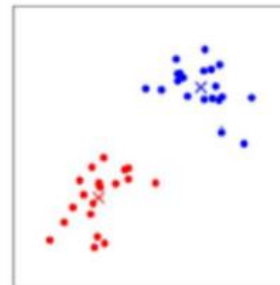
(c)



(d)



(e)



(f)

K-Means

Hierarchical
Clustering

K평균 군집화
K-means

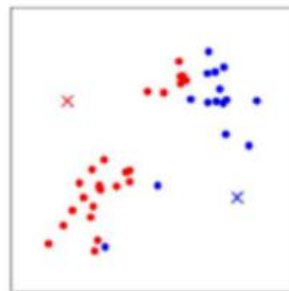
DBSCAN



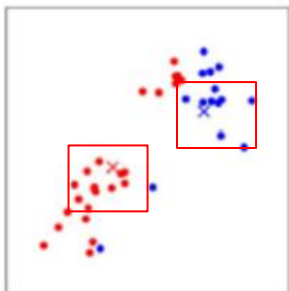
(a)



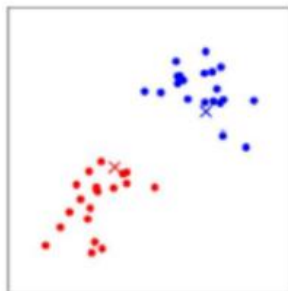
(b)



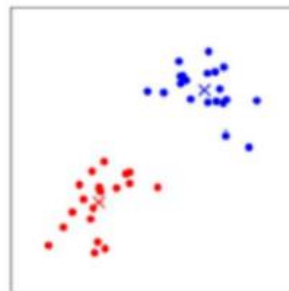
(c)



(d)



(e)



(f)

K-Means

Hierarchical
Clustering

K평균 군집화
K-means

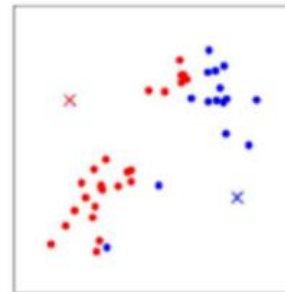
DBSCAN



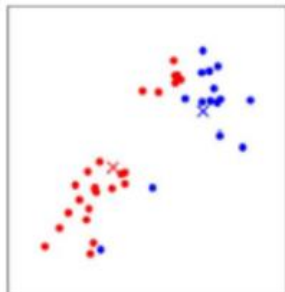
(a)



(b)



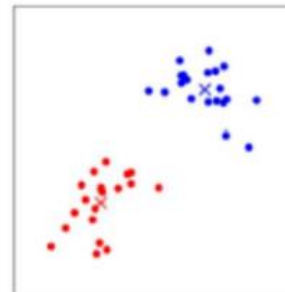
(c)



(d)



(e)



(f)

K-Means

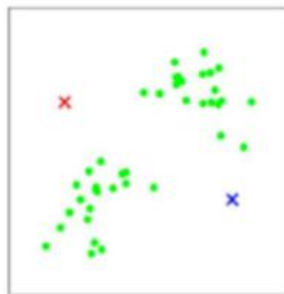
Hierarchical
Clustering

K평균 군집화
K-means

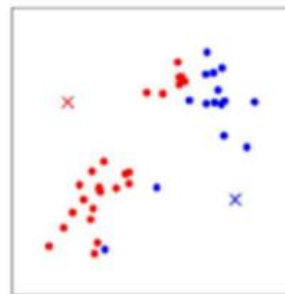
DBSCAN



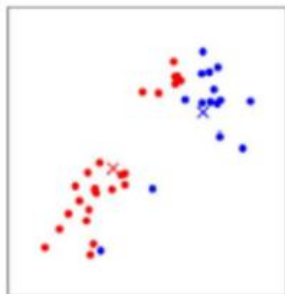
(a)



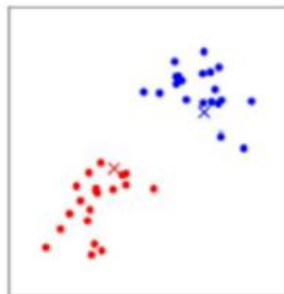
(b)



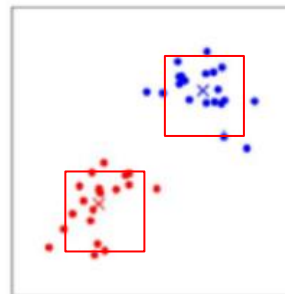
(c)



(d)



(e)



(f)

K-Means

Hierarchical
Clustering

K평균 군집화
K-means

DBSCAN

장점

- 알고리즘이 쉽고 간결함
- 대용량 데이터에도 활용가능함

단점


- Feature의 개수가 많은 경우 군집화 정확도가 떨어짐
- 반복 횟수가 많을 경우 수행 시간이 오래 걸림
- 몇 개의 군집을 선택해야 할지 정하기가 어려움
- 이상치(outlier)데이터에 취약함

K-Means

Hierarchical
Clustering

K평균 군집화
K-means

DBSCAN

위젯	설명	입력	출력
 k-Means	k-평균 군집화 알고리즘을 사용하여 항목을 그룹화한다.	Data	Data, Centroids

- k-Means 위젯은 데이터에 k-Means 클러스터링 알고리즘을 적용하고 클러스터 인덱스가 클래스 속성으로 사용되는 새 데이터 세트를 출력
- 원래 클래스 속성이 있는 경우 메타 데이터 속성으로 이동

K-Means

Hierarchical
Clustering

K평균 군집화
K-means

DBSCAN

① Number of Clusters

군집 수를 선택한다.

Fixed: 알고리즘이 지정된 수의 클러스터로 데이터를 클러스터링한다.

From~ to~: 선택한 클러스터 범위에 대한 클러스터링 점수를 보여준다.

② Preprocessing

열 정규화를 실시한다.

③ Initialization

초기화 방법(알고리즘이 클러스터링을 시작하는 방법)을 선택한다.

k-Means++: 첫 번째 중심은 랜덤하게 선택되고, 그 이후는 가장 가까운 중심으로부터의 거리 제곱에 비례하는 확률로 나머지 점으로부터 선택된다.

Random initialization: 처음에는 클러스터가 랜덤으로 할당되고 이후 반복으로 업데이트된다.

Re-run: 알고리즘이 랜덤 초기 위치에서 실행되는 횟수, 클러스터 내 제곱합이 가장 낮은 결과가 사용된다.


maximal iterations: 각 알고리즘 실행 내 최대 반복 횟수를 설정한다.

K-Means

Hierarchical
Clustering

K평균 군집화
K-means

DBSCAN

위젯	설명	입력	출력
 Silhouette Plot	데이터 군집 내의 일관성을 그래픽으로 나타낸다.	Data	Selected Data, Data

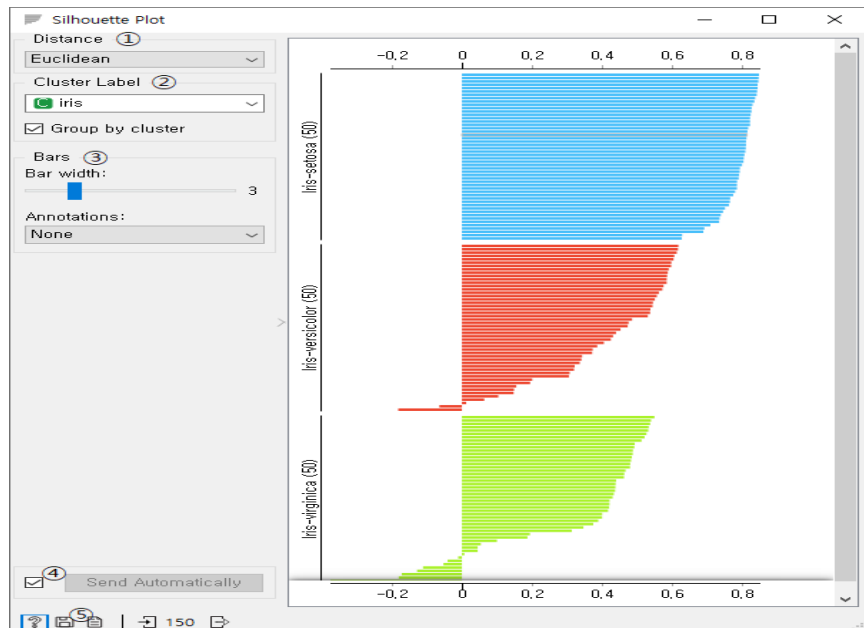
- Silhouette Plot 위젯은 데이터 군집 내의 일관성을 그래픽으로 표현하고 클러스터 품질을 시각적으로 평가하는 방법을 사용자에게 제공
- 실루엣 점수는 물체가 다른 군집과 비교했을 때 자신의 군집과 얼마나 유사한지를 나타내는 척도로 실루엣 플롯의 생성에 결정적
- 실루엣 점수가 1에 가까우면 데이터 인스턴스가 클러스터의 중심에 가깝고 실루엣 점수가 0에 가까운 인스턴스가 두 클러스터 사이의 경계에 있음을 알 수 있음.

K-Means

Hierarchical
Clustering

K평균 군집화
K-means

DBSCAN



거리 메트릭스를 선택한다.

① Distance

Euclidean

두 점 사이의 직선 거리

Manhattan

모든 속성에 대한 절대 차이의 합계

Cosine

1-두 벡터 사이의 각도의 코사인

② Cluster Label

클러스터 레이블을 선택한다. 인스턴스를 클러스터별로 그룹화할지 여부를 결정할 수 있다.

③ Bars

막대 너비를 선택하고 실루엣 그림에 주석(annotations)을 달 수 있다.

④ Send
Automatically

데이터를 자동으로 반영한다.

⑤ 도움말

이미지를 컴퓨터에 .svg 또는 .png 형식으로 저장하거나 보고서를 작성한다.

K-Means

Hierarchical
Clustering

K평균 군집화
K-means

DBSCAN



k-Means ? X

Number of Clusters

☒ Fixed:

☐ From to

Preprocessing

☒ Normalize columns

Initialization

Initialize with KMeans++

Re-runs:

Maximum iterations:

Apply Automatically

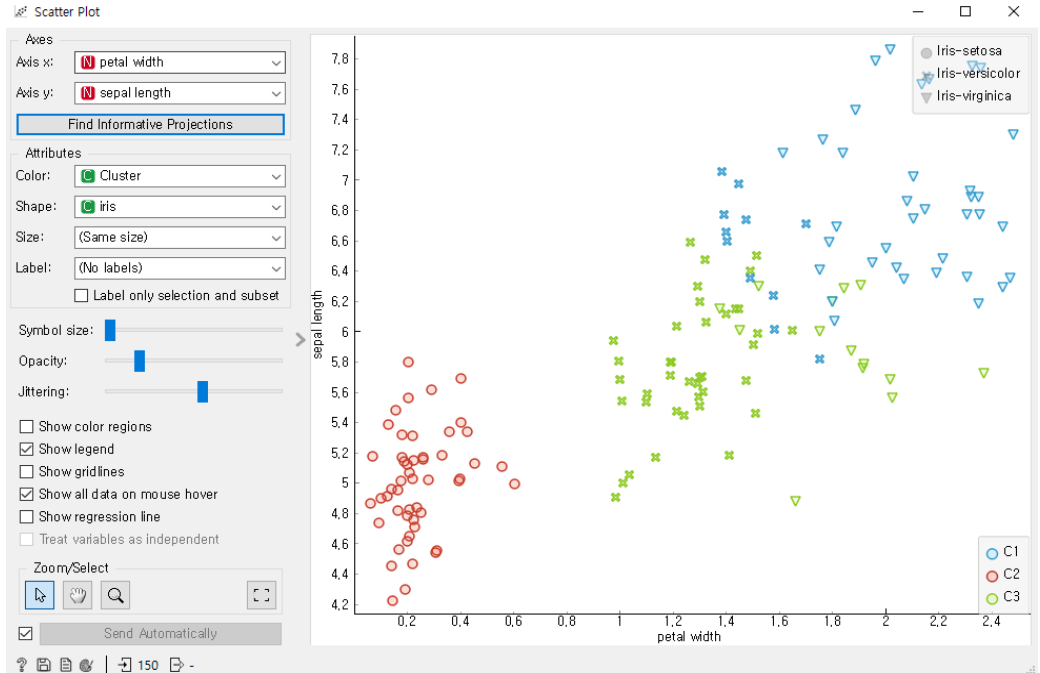
? | 150 150

K-Means

Hierarchical
Clustering

K평균 군집화
K-means

DBSCAN

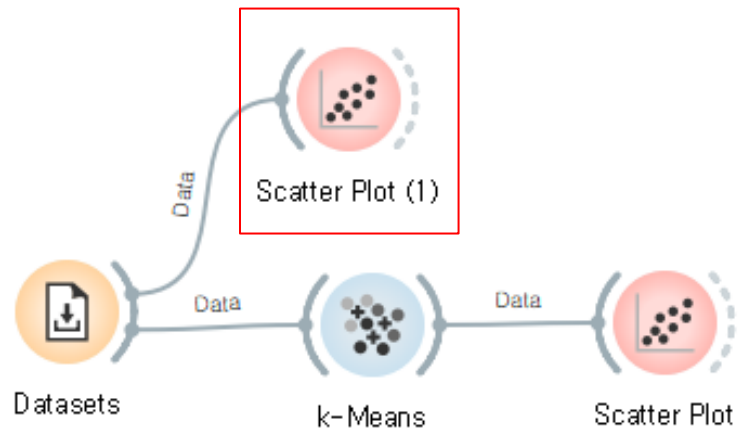


K-Means

Hierarchical
Clustering

K평균 군집화
K-means

DBSCAN

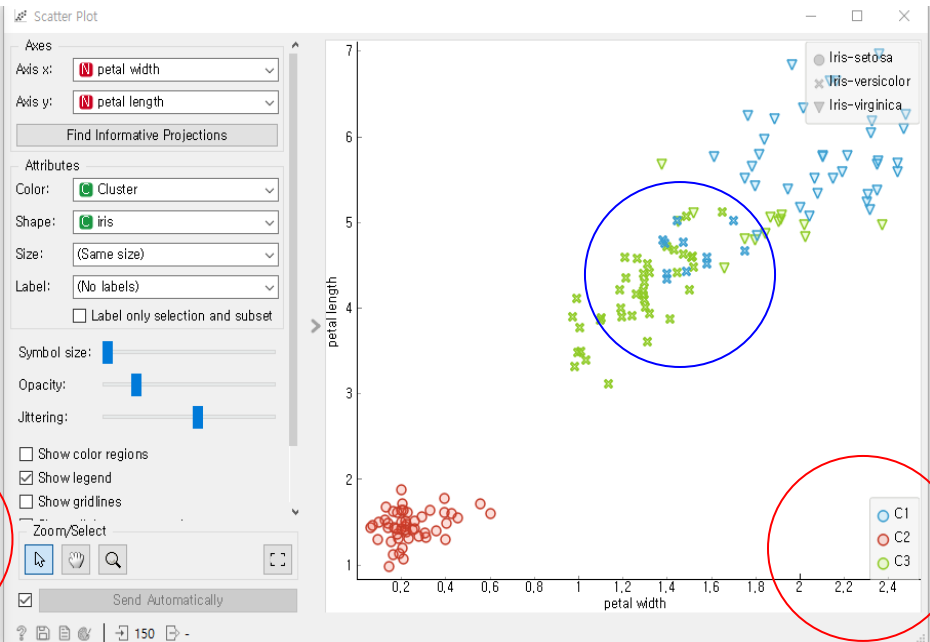
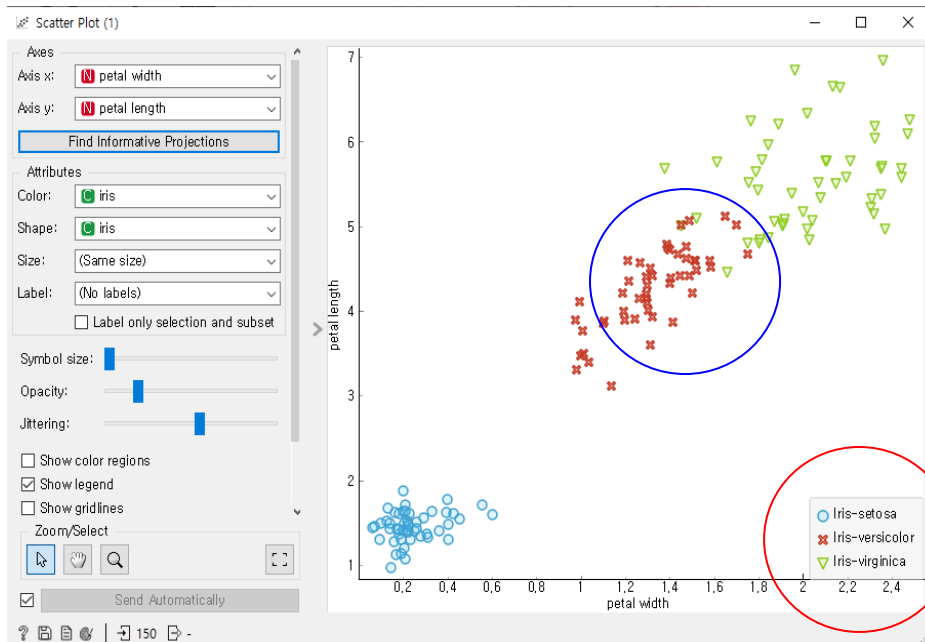


K-Means

Hierarchical
Clustering

K평균 군집화
K-means

DBSCAN

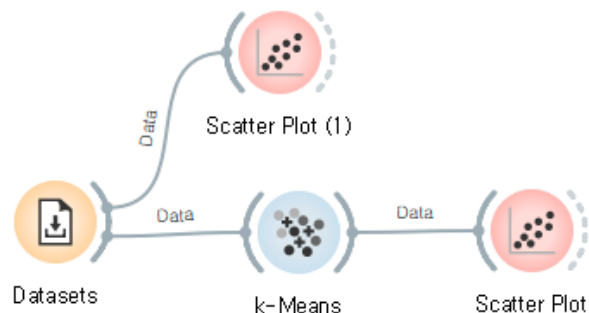


K-Means

Hierarchical
Clustering

K평균 군집화
K-means

DBSCAN



k-Means

Number of Clusters

☐ Fixed: 3

☒ From 2 to 8

Preprocessing

☒ Normalize columns

Initialization

Initialize with KMeans++

Re-runs: 10

Maximum iterations: 300

Apply Automatically

Silhouette Scores

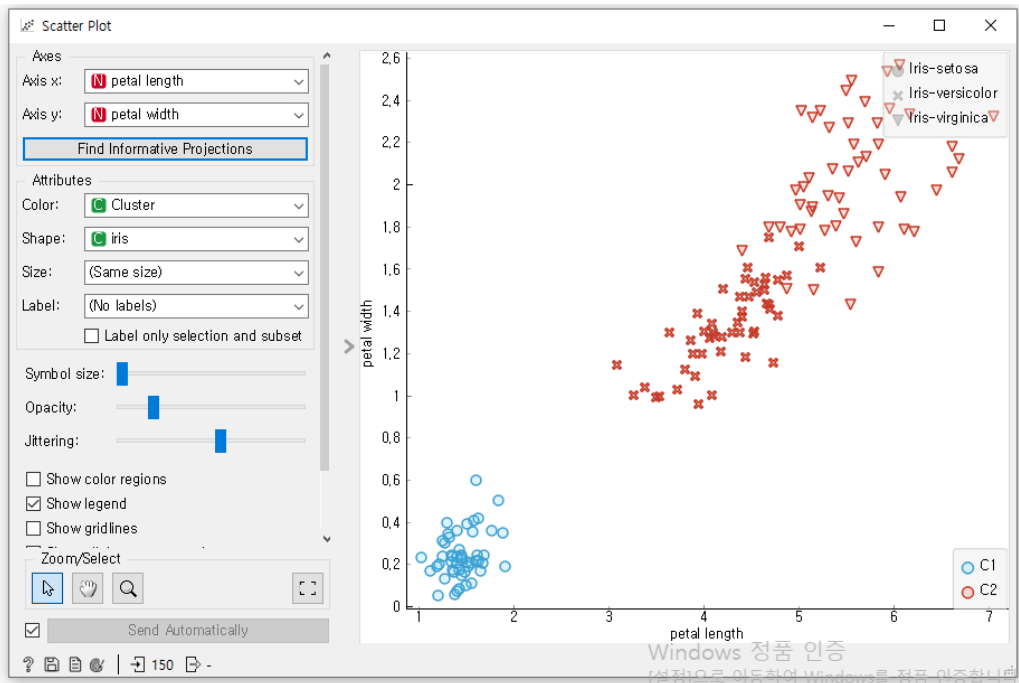
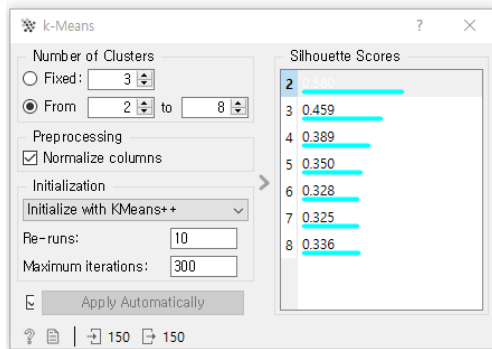
2	0.580
3	0.459
4	0.389
5	0.350
6	0.328
7	0.325
8	0.336

K-Means

Hierarchical
Clustering

K평균 군집화
K-means

DBSCAN

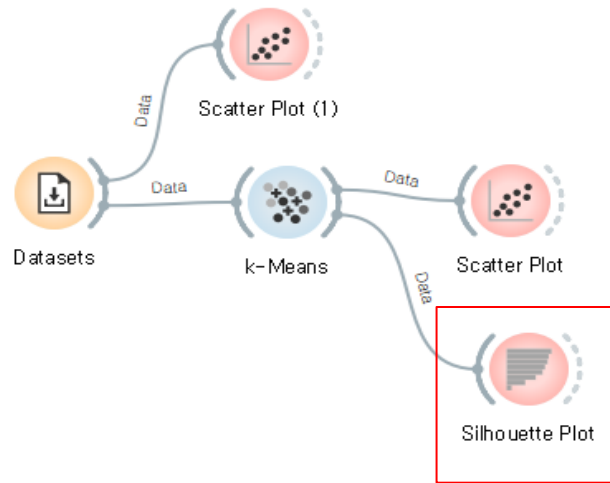


K-Means

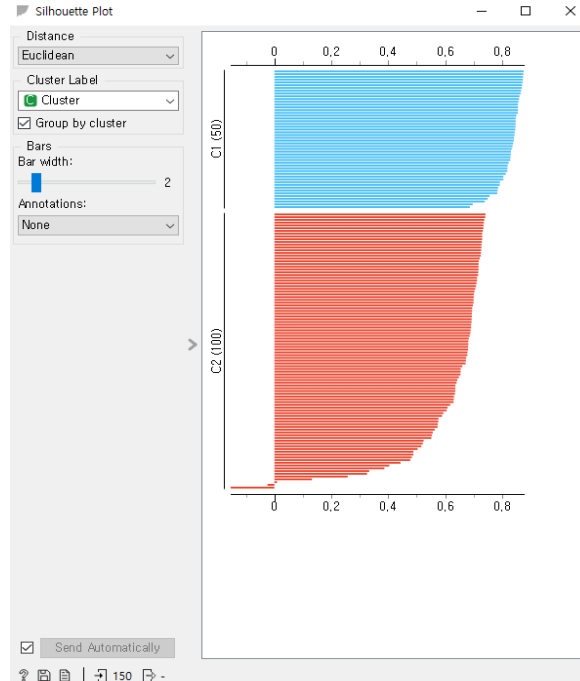
Hierarchical
Clustering

K평균 군집화
K-means

DBSCAN



실루엣스코어는 군집화가 잘 되었는지 평가

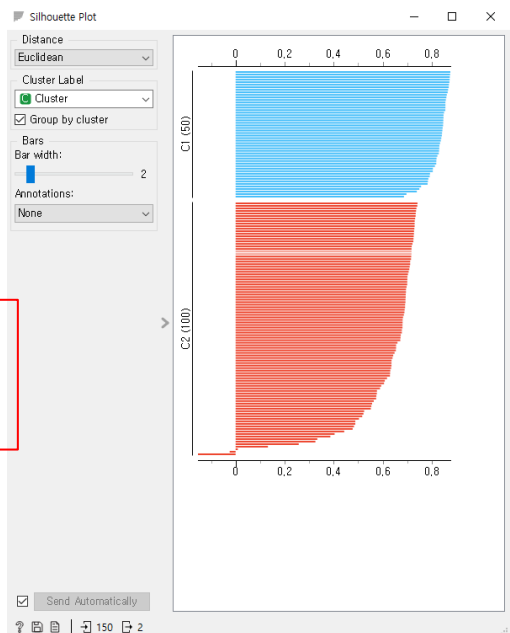
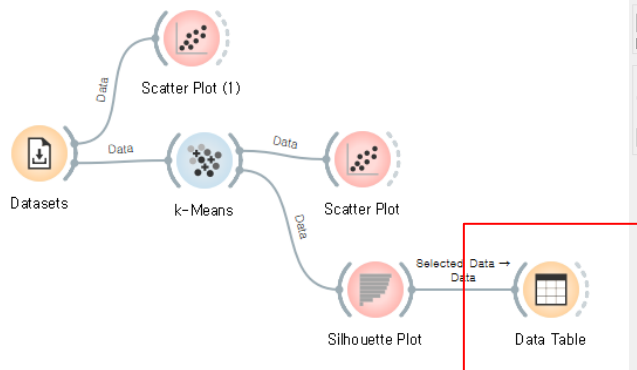


K-Means

Hierarchical Clustering

K평균 군집화
K-means

DBSCAN



Data Table

Info

- 2 instances (no missing data)
- 4 features
- Target with 3 values
- meta attributes

Variables

- ☒ Show variable labels (# present)
- ☒ Visualize numeric values
- ☒ Color by instance classes

Selection

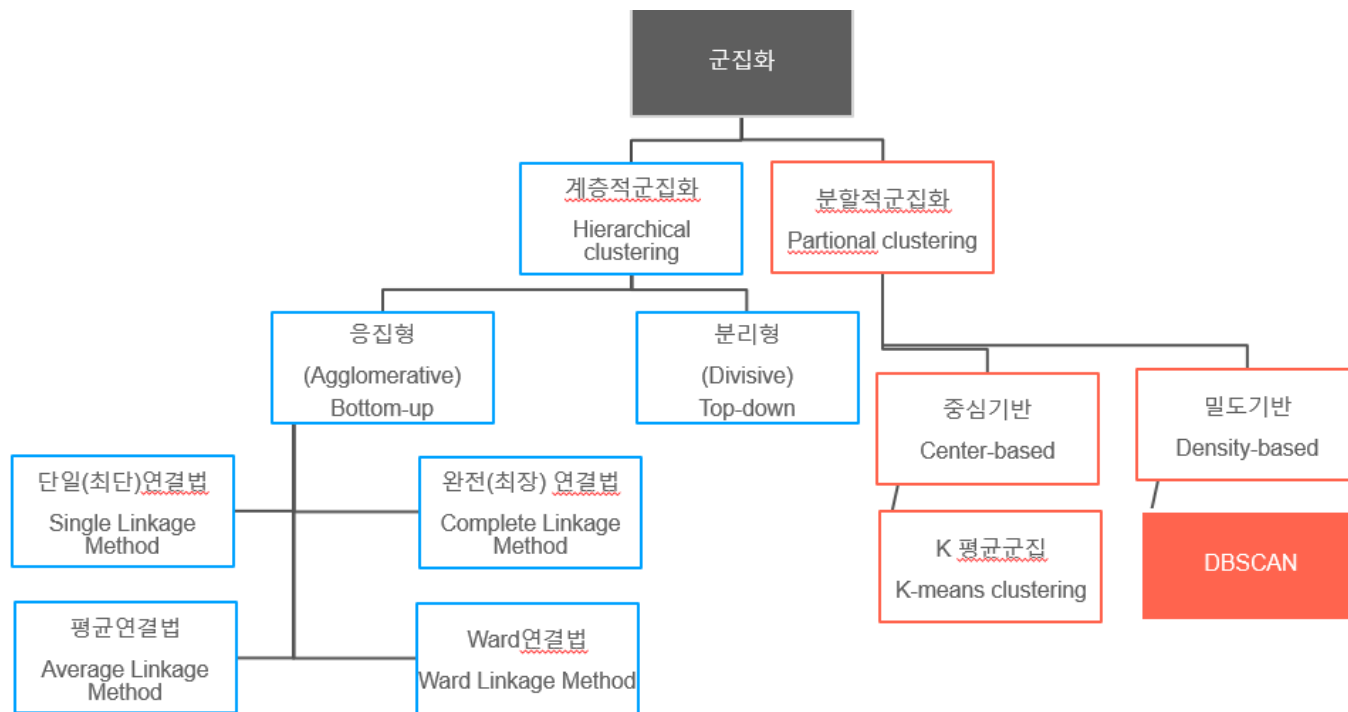
- ☒ Select full rows

Restore Original Order

☒ Send Automatically

	iris	Cluster	Silhouette	silhouette (Cluster)	sepal length
1	iris-virginica	C2	0.675896	0.717611	6.0
2	iris-virginica	C2	0.678626	0.717629	6.9

군집화 모델

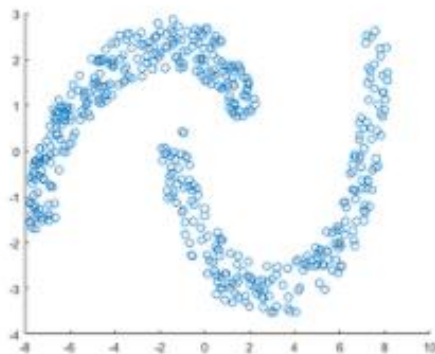


밀도기반 클러스터링(DBSCAN)

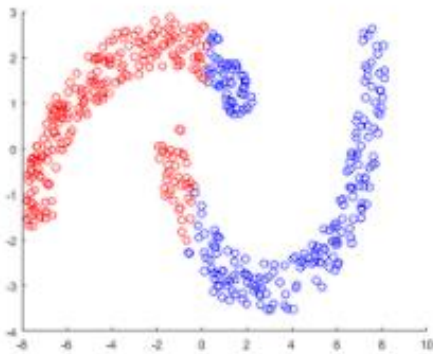
Hierarchical
Clustering

K평균 군집화
K-means

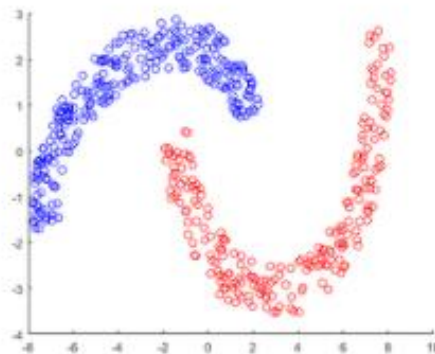
DBSCAN



(a) 원본 데이터



(b) k-means clustering의 결과



(c) DBSCAN의 결과

밀도기반 클러스터링(DBSCAN)

Hierarchical
Clustering

K평균 군집화
K-means

DBSCAN

DBSCAN은 각각의 데이터들에 대해 이웃한 데이터와의 밀도를
계산하면서 불특정한 모양의 클러스터를 생성한다.

*DBSCAN(Density-based spatial clustering of applications with noise)

밀도 기반의 군집화는 점이 촘촘하게 몰려 있어서 “단위 면적당 데이터 점들의 개수인
밀도가 높은 부분을 군집화” 하는 방식

밀도기반 클러스터링(DBSCAN)

Hierarchical
Clustering

K평균 군집화
K-means

DBSCAN

장점

- 클러스터 수를 지정할 필요가 없으며 알고리즘이 자동으로 클러스터 수를 찾음
- 원 모양 뿐만 아니라 불특정한 모양의 클러스터도 찾음
- 클러스터링과 동시에 노이즈데이터도 분류할 수 있어 outlier에 의한 성능 저하는 완화할 수 있음

단점


- 데이터가 입력되는 순서에 따라 클러스터링 결과가 변함
- 알고리즘이 이용하는 거리 측정 방법에 따라 클러스터링 결과가 변함
- 데이터의 특성을 모를 경우에는 알고리즘의 적절한 hyper-parameter를 설정하기가 어려움

밀도기반 클러스터링(DBSCAN)

Hierarchical
Clustering

K평균 군집화
K-means

DBSCAN

위젯	설명	입력	출력
 DBSCAN	DBSCAN 클러스터링 알고리즘을 사용하여 항목을 그룹화한다.	Data	Data

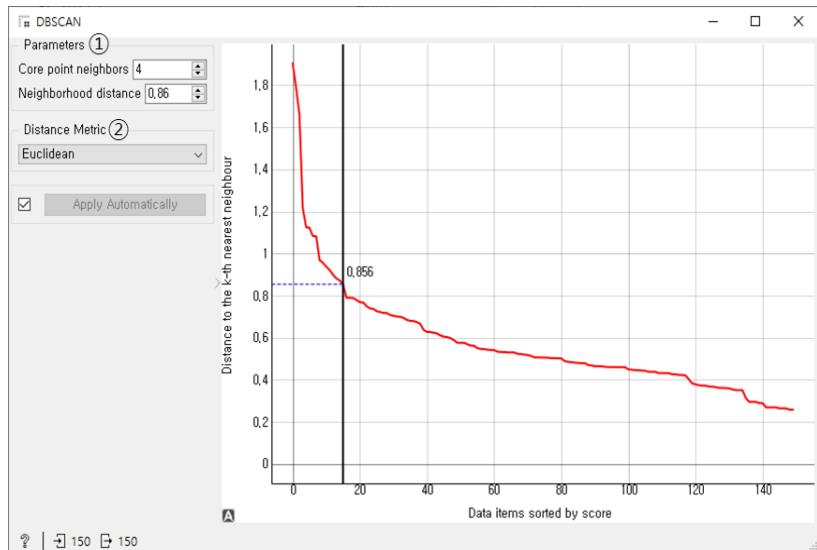
- DBSCAN 위젯은 **DBSCAN 클러스터링 알고리즘을 데이터에 적용**하고 클러스터 인덱스를 메타데이터 속성으로 사용하여 새 데이터 세트를 출력
- 또한 위젯은 k번째 가장 가까운 인접 거리가 있는 정렬된 그래프도 표시
- k값은 Core point neighbors로 설정.
- 이를 통해 사용자는 Neighborhood 거리 설정에 이상적인 선택을 할 수 있음.

밀도기반 클러스터링(DBSCAN)

Hierarchical
Clustering

K평균 군집화
K-means

DBSCAN



① Parameters

클러스터에 대한 최소 코어 인접 지역 수와 최대 인접 지역 거리를 설정한다.

② Distance Metric

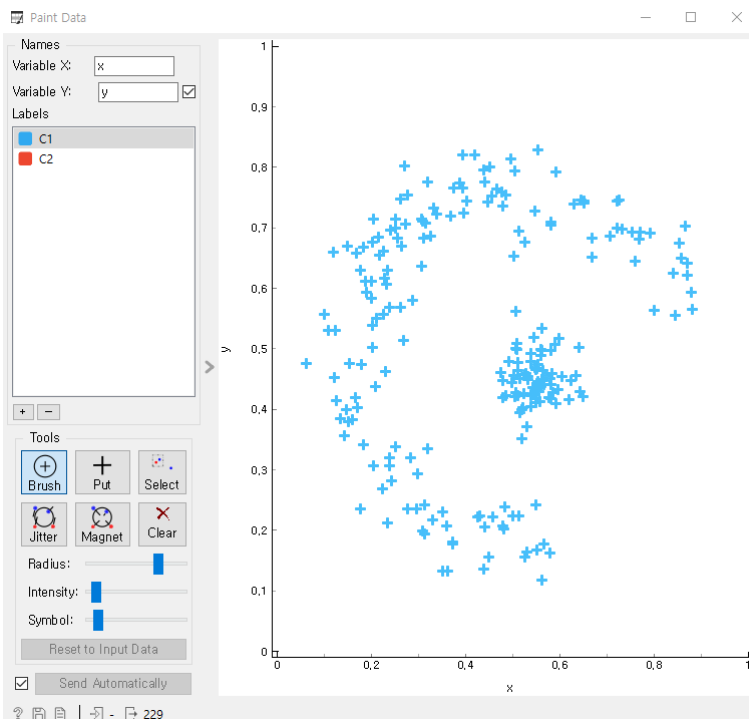
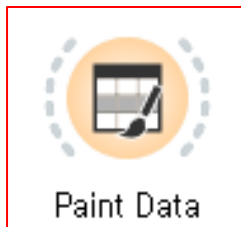
항목을 그룹화하는 데 사용되는 거리 메트릭을 설정한다.

밀도기반 클러스터링(DBSCAN)

Hierarchical
Clustering

K평균 군집화
K-means

DBSCAN

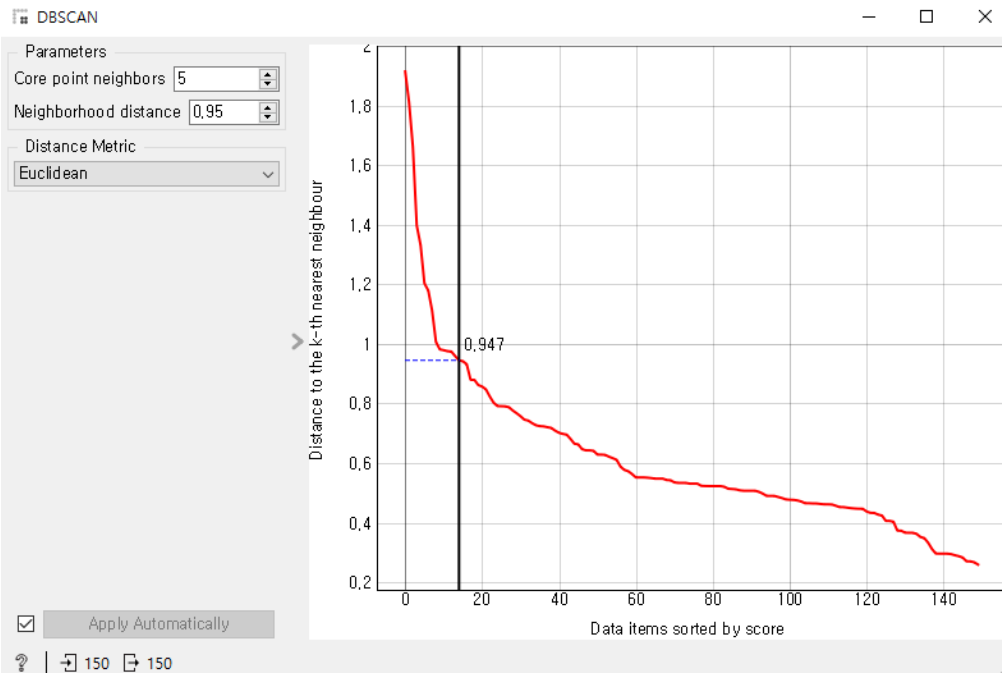
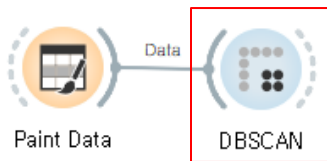


밀도기반 클러스터링(DBSCAN)

Hierarchical
Clustering

K평균 군집화
K-means

DBSCAN

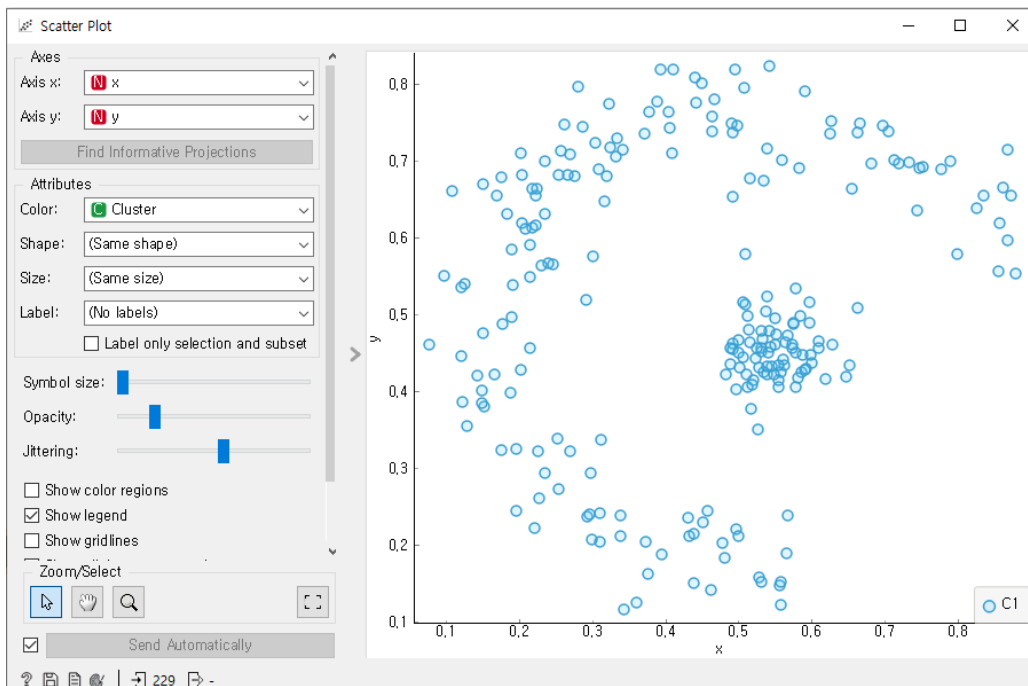
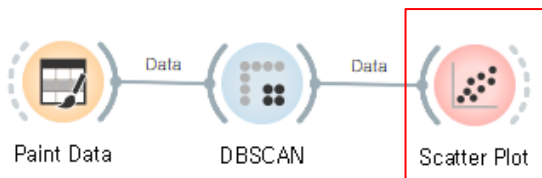


밀도기반 클러스터링(DBSCAN)

Hierarchical
Clustering

K평균 군집화
K-means

DBSCAN

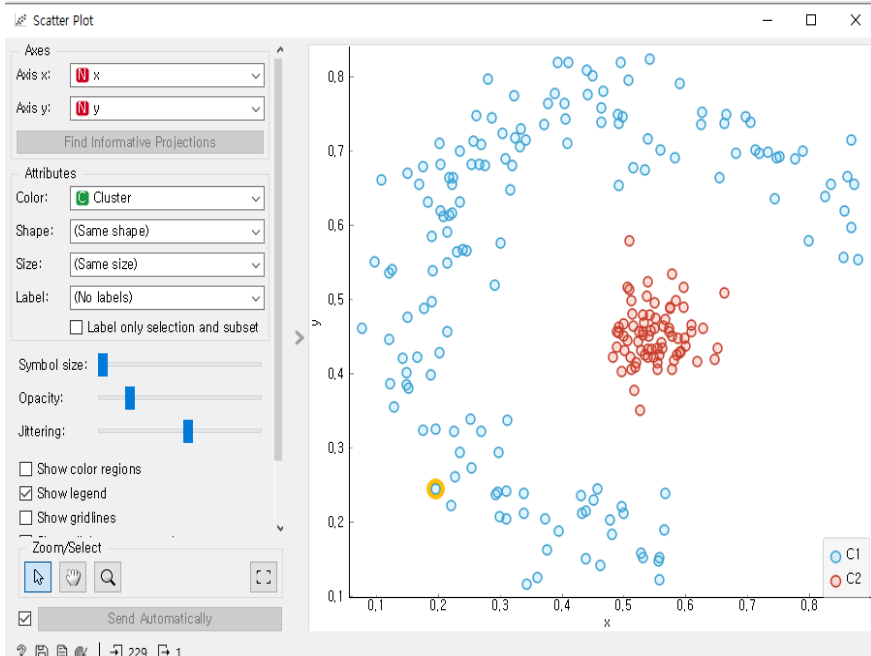
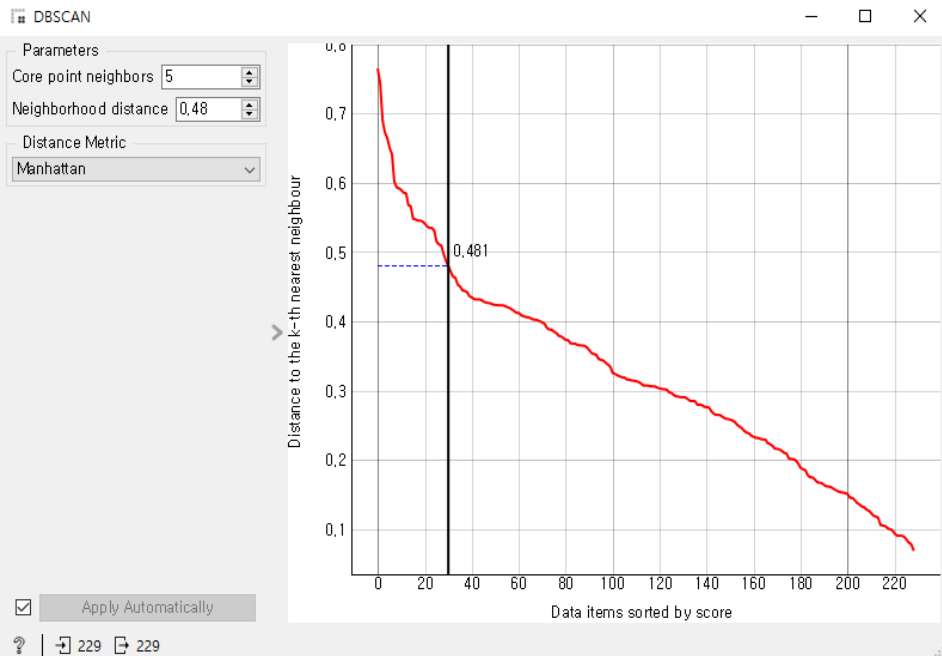


밀도기반 클러스터링(DBSCAN)

Hierarchical
Clustering

K평균 군집화
K-means

DBSCAN

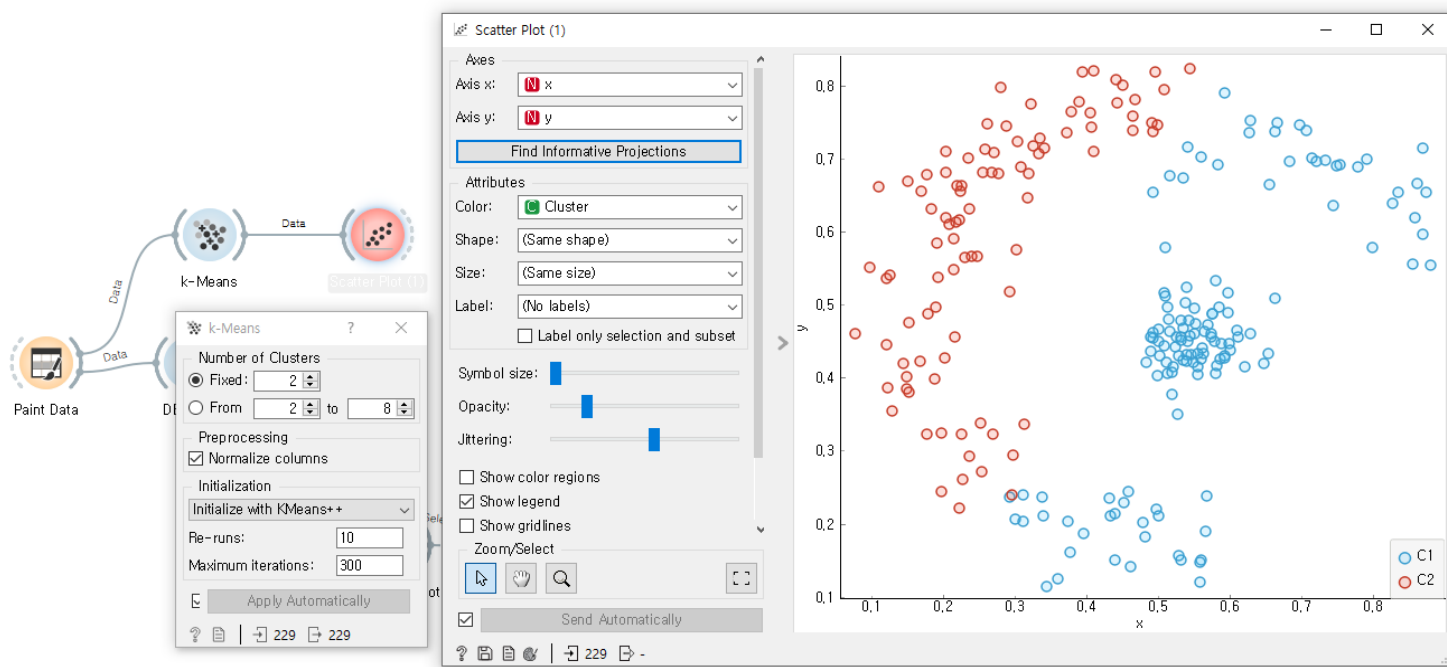


밀도기반 클러스터링(DBSCAN)

Hierarchical
Clustering

K평균 군집화
K-means

DBSCAN

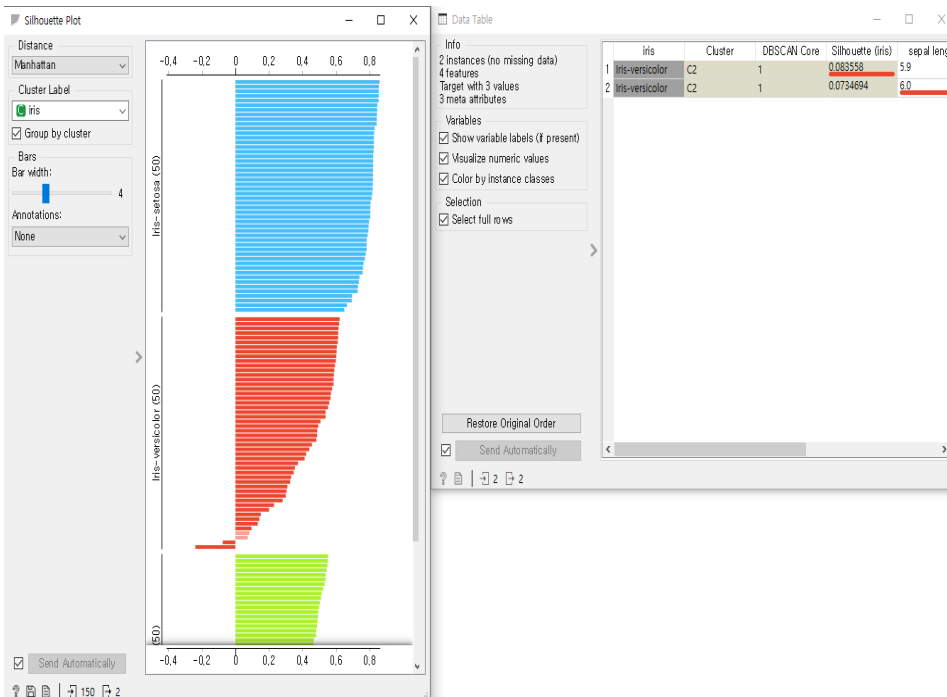
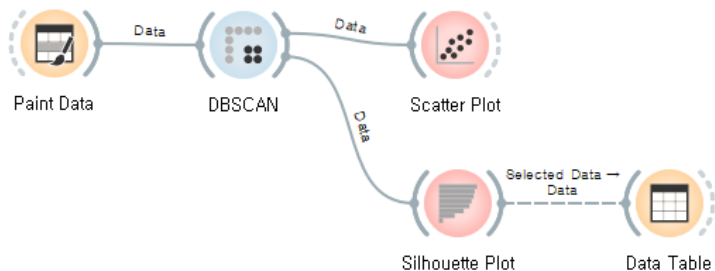


밀도기반 클러스터링(DBSCAN)

Hierarchical
Clustering

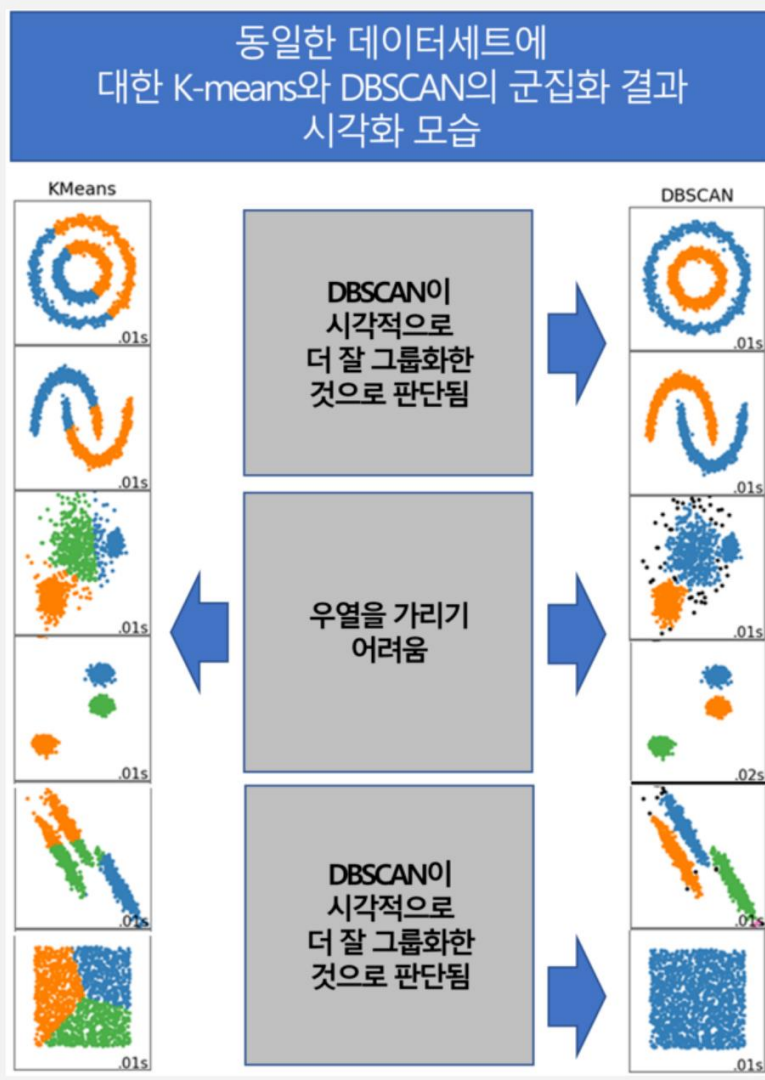
K평균 군집화
K-means

DBSCAN



K-Means vs. DBSCAN

Source: https://scikit-learn.org/stable/auto_examples/cluster/plot_cluster_comparison.html#sphx-glr-auto-examples-cluster-plot-cluster-comparison-py



질문 있나요?

hsryu13@hongik.ac.kr

