DBSCAN (Density-Based Spatial Clustering of Applications with Noise) - це метод кластеризації, який можна використовувати для розділення наборів даних на кластери на основі їх густоти. Основна ідея методу полягає в тому, що кластери формуються з областей, де є висока густота точок, і розділяються зонами з низькою густотою.

DBSCAN працює з двома параметрами: радіусом (eps) та мінімальною кількістю точок (minPts), які повинні бути знаходитися в радіусі eps, щоб розглядатися як частина кластера.

Алгоритм на вхід приймає параметри (eps, minPts) та точки, які треба розбити на кластери. Вихідні данні частіше всього видаються у вигляді матриці типів, де кожній точці видаться номер кластеру до якого вона входить.

Допоміжні конструкції: часто використовується матриця відстаней, де розрахована відстань між кожними точками.

Алгоритм DBSCAN працює наступним чином:

1. Вибираємо довільну точку з датасету, якщо ця точка має достатньо сусідів (кількість точок, які знаходяться в радіусі eps більше або дорівнює minPts), то вона стає частиною нового кластера. Якщо точка не має достатньо сусідів, то вона вважається шумом.
2. Якщо, точка має достатньо сусідів то створюється кластер і її сусіди додаються до клстера та також перевіряються, якщо вони проходять перевірку, то включається перевірка для їх сусідів.
3. Крок 2 повторюється, поки всі сусіди кластера не будуть перевірені. Далі знову повторюємо перші два кроки доки всі точки не будуть відвідані.

Додаткові правила пункту 2:

* Якщо в результаті перевірки всіх сусідів виявляється, що точка не має достатньо сусідів, то вона вважається шумом.
* Якщо під час перевірки точок виявляється, що декілька кластерів перетинаються, то вони об'єднуються в один кластер.

DBSCAN має кілька переваг порівняно з іншими методами кластеризації. Зокрема, він може відокремлювати кластери з різною густотою, а також здатний розрізняти шум від кластерів. Крім того, DBSCAN не вимагає заздалегідь визначеного числа кластерів, що дозволяє йому знаходити кластери автоматично.

Однак, DBSCAN має декілька недоліків. Наприклад, якщо в датасеті присутні кластери з різною густотою, то задані значення eps та minPts можуть бути не оптимальними для всіх кластерів. Крім того, алгоритм може мати проблеми з кластеризацією великих наборів даних, оскільки він має складність O(n^2).

У практичному застосуванні DBSCAN можна використовувати для різних завдань, таких як кластеризація зображень, виявлення аномалій в даних або виявлення груп користувачів в соціальних мережах.

У висновку, DBSCAN - це потужний метод кластеризації, який здатен автоматично визначати кластери без заздалегідь визначеного числа кластерів. Однак, для досягнення оптимальних результатів, необхідно коректно підібрати значення параметрів eps та minPts.