DBSCAN Algoritması: Teori, Avantajlar, ve Kullanım

Atıl Samancıoğlu

1 Giriş

DBSCAN (**Density-Based Spatial Clustering of Applications with Noise**), yoğunluk tabanlı bir kümeleme algoritmasıdır. K-means gibi küme sayısını baştan bilmek zorunda değildir ve veri setindeki **gürültü (outlier)** verileri doğal olarak tespit edebilir.

2 Temel Kavramlar

DBSCAN üç tür veri noktası tanımlar:

- **Core Point (Çekirdek Nokta):** Etrafında yeterli sayıda veri noktası varsa (bir yoğunluk oluşturuyorsa).
- **Border Point (Sınır Noktası):** Bir core point'in komşuluğundadır ama kendi başına yoğunluk oluşturmaz.
- **Outlier (Gürültü):** Hiçbir kümenin parçası olmayan noktadır.

İki önemli parametresi vardır:

- ε (Epsilon): Nokta etrafındaki yarıçap (komşuluk mesafesi).
- `minPts`: Bir noktanın core olabilmesi için gerekli minimum komşu sayısı.

3 Matematiksel Tanım

Epsilon-komşuluk: Bir nokta p için, ε yarıçapındaki komşuluk aşağıdaki gibi tanımlanır:

$$N_\varepsilon(p) = \{q \in D \mid \text{dist}(p, q) \leq \varepsilon\}$$

Core Point: Eğer $|N_\varepsilon(p)| \geq$ `minPts` ise p bir core point'tir.
Direct Density Reachable: q noktası, p'den doğrudan yoğunlukla erişilebiliyorsa:

$$q \in N_\varepsilon(p) \quad \text{ve} \quad p \text{ bir core point ise}$$

Density Reachable: Bir dizi doğrudan erişilebilir noktalar zinciri varsa q noktası p'den yoğunlukla erişilebilirdir.
Density Connected: Eğer p ve q noktaları bir üçüncü nokta o üzerinden erişilebiliyorsa, p ve q yoğunlukla bağlantılıdır.

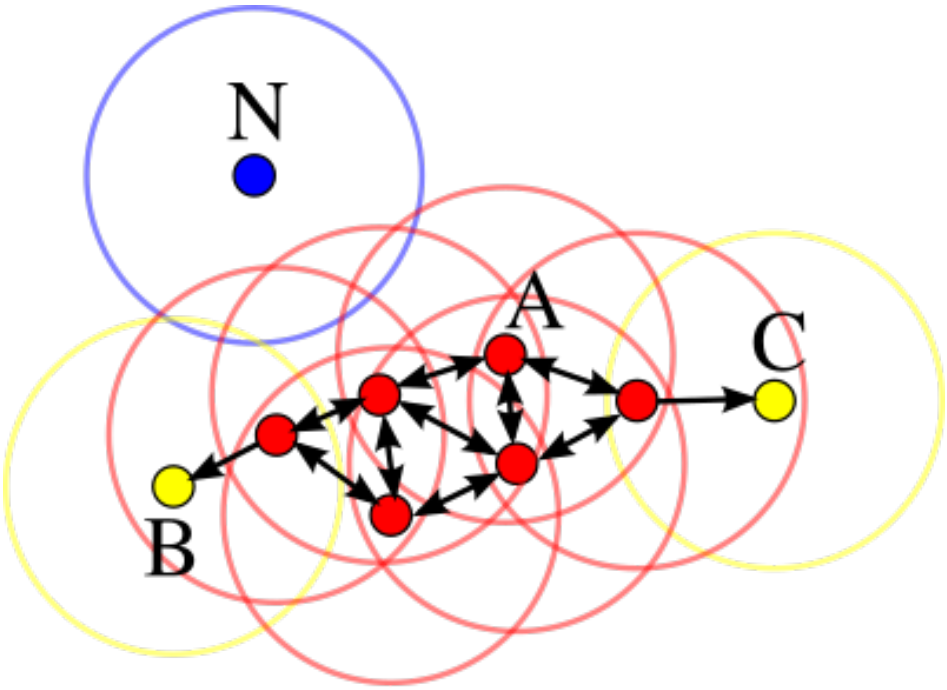

Figure 1: Kaynak: wikipedia

4 DBSCAN Algoritması Adımları

1. Tüm noktalar ziyaret edilmemiş olarak işaretlenir.
2. Herhangi bir nokta p seçilir.
3. $N_\varepsilon(p)$ hesaplanır.
4. Eğer $|N_\varepsilon(p)| < \texttt{minPts}$ ise:
   - p, başka bir core point'in ε-komşuluğu içindeyse **border point** olarak etiketlenir.
   - p, hiçbir core point'e komşu değilse **noise (outlier)** olarak etiketlenir.
5. Aksi takdirde p bir core point olarak kabul edilir ve yeni bir küme başlatılır.
6. Tüm $N_\varepsilon(p)$ noktaları kümeye dahil edilir. Eğer bunlar da core point ise onların komşuları da kümeye alınır.
7. Tüm noktalar ziyaret edilene kadar işlem devam eder.

5 Avantajlar ve Dezavantajlar

Avantajlar

- Küme sayısı önceden belirtilmek zorunda değildir.
- Gürültü verilerini otomatik olarak tespit eder.
- **Non-lineer şekilli kümeleri** başarıyla ayırabilir.
- K-means'e göre daha esnek ve gerçek hayata daha uygundur.

Dezavantajlar

- Yoğunluğu değişken veri setlerinde başarısız olabilir.
- Parametre seçimi zordur (özellikle ε).
- Küçük kümeler büyük kümeler tarafından yutulabilir.
- Sınırlı ölçeklenebilirlik (çok büyük veri setlerinde yavaş olabilir).

6 K-means ve DBSCAN Karşılaştırması

Özellik	**K-means**	**DBSCAN**
Küme Sayısı	Belirtilmeli	Otomatik tespit
Outlier Tespiti	Yapamaz	Doğrudan ayırır
Küme Şekli	Küresel	Serbest/Non-lineer
Yoğunluk Duyarlılığı	Duyarsız	Duyarlı

Table 1: K-means vs. DBSCAN Karşılaştırması

7 Ne Zaman DBSCAN Kullanmalı?

- Veri dağılımı yoğunluk bazlıysa,
- Küme sayısı önceden bilinmiyorsa,
- Gürültü verisi yüksekse,
- Non-lineer yapılar varsa.

8 HDBSCAN ve DBSCAN Arasındaki Farklar

HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise), DBSCAN algoritmasının bir genellemesidir. HDBSCAN, sabit bir ε değeri yerine farklı yoğunluk seviyelerinde kümeleme yapar ve sonuçları bir **yoğunluk temelli hiyerarşi** olarak üretir. Bu hiyerarşiden en istikrarlı kümeler seçilir.

HDBSCAN'in Temel Özellikleri

- **Parametre sayısı daha azdır**: HDBSCAN için yalnızca `min_cluster_size` parametresi gereklidir, ε gibi hassas bir parametre gerekmez.
- **Yoğunluk çeşitliliğine duyarlıdır**: Farklı yoğunluktaki bölgeleri daha doğru bir şekilde kümeleyebilir.
- **Hiyerarşik çıktı üretir**: Küme yapısı farklı seviyelerde detaylandırılabilir, böylece kullanıcı farklı ayrıntı seviyelerinde analiz yapabilir.
- **Noise (gürültü) tanımı korunur**: DBSCAN gibi HDBSCAN de outlier noktaları tanıyıp ayırabilir.

DBSCAN	HDBSCAN
Sabit bir ε yarıçapı kullanır.	Farklı yoğunluk seviyelerinde analiz yapar, ε gerekmez.
Tüm kümeler aynı yoğunluk eşiğine sahiptir.	Küme yoğunlukları esnek şekilde değişebilir.
Kümeler sabittir, hiyerarşi içermez.	Kümeler hiyerarşik yapıda çıkarılır.
Parametre seçimi zordur (özellikle ε).	Daha az parametre ile daha istikrarlı sonuçlar elde edilir.
Hızlıdır ancak düşük yoğunluklu kümelerde başarısız olabilir.	Daha hesaplama yoğun ama daha esnek ve doğru sonuçlar üretir.

Table 2: DBSCAN ve HDBSCAN Arasındaki Farklar

DBSCAN ile Farkları

Matematiksel Açıdan HDBSCAN ile DBSCAN Arasındaki Farklar

DBSCAN algoritması temel olarak sabit bir ε yarıçapı etrafındaki yoğunluğu ölçerek kümeler oluşturur. Bu sabit mesafe nedeniyle tüm kümeler aynı yoğunluk varsayımıyla oluşturulur.

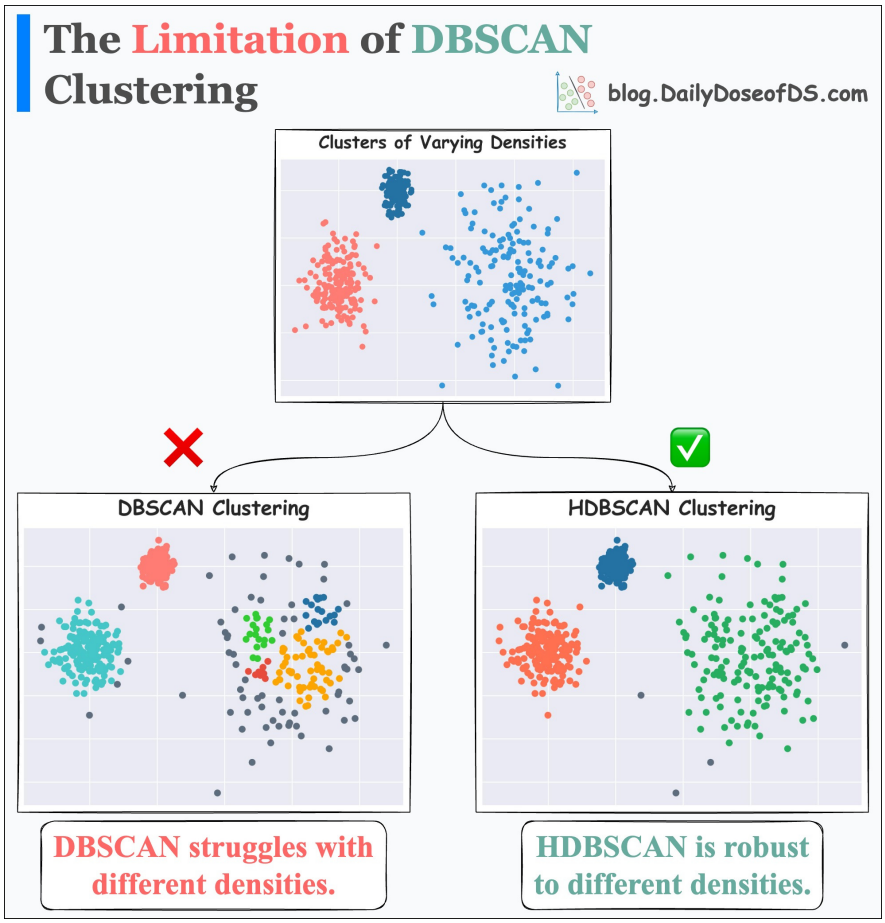

Figure 2: Kaynak: daily dose of ds

HDBSCAN ise bu sabit mesafe varsayımını ortadan kaldırır. Onun yerine şu matematiksel bileşenlerle

çalışır:

1. **Mutual Reachability Distance (Ortak Erişim Mesafesi):** İki nokta arasındaki mesafe sadece doğrudan aralarındaki mesafe ile değil, aynı zamanda lokal yoğunluk farkı ile de ölçülür:

$$d_{\text{mreach-k}}(a, b) = \max\left(\text{core}_k(a), \text{core}_k(b), d(a, b)\right)$$

   Burada $\text{core}_k(a)$, nokta a için k en yakın komşusuna olan mesafeyi ifade eder. Bu formül, düşük yoğunluklu bölgelerdeki noktaların birbirine olan mesafesini daha büyük gösterir ve böylece bu noktaların daha zor birleşmesini sağlar.

2. **Minimum Spanning Tree (MST) Oluşturma:** Tüm noktaların mutual reachability mesafelerine göre bir minimum spanning tree (MST) oluşturulur. Bu ağ, tüm noktaları birbirine en kısa yoğunluk-bilinçli yollarla bağlar.

3. **Yoğunluk Bazlı Kümeleme:** MST üzerinde kenar uzunluklarına göre artan şekilde kesilerek bir kümeleme hiyerarşisi oluşturulur. Her bir kesim, daha seyrek bağlantıları ayırır ve yoğun kümeleri ortaya çıkarır.

4. **Stabiliteye Dayalı Küme Seçimi:** HDBSCAN, küme hiyerarşisinden en **stabil** kümeleri otomatik olarak seçer. Stabilite, bir kümenin var olabildiği yoğunluk aralığının uzunluğuna bağlıdır:

$$\text{Stabilite} = \sum_{x \in C} (\lambda_{\text{birth}}(x) - \lambda_{\text{death}}(x))$$

   Burada $\lambda = \frac{1}{\text{reachability}}$ olarak tanımlanır. Daha uzun ömürlü (yoğunluk değişimine daha dayanıklı) kümeler tercih edilir.

Özetle Matematiksel Farklar

- DBSCAN sabit ε kullanır, HDBSCAN değişken yoğunlukları destekleyen **mutual reachability** mesafesi kullanır.
- DBSCAN sadece kümeleri üretir, HDBSCAN önce bir **hiyerarşi** üretir, sonra stabil kümeleri seçer.
- DBSCAN'de **hard threshold** ile bölünür, HDBSCAN'de **minimum spanning tree** ve stabilite kavramı kullanılır.
- HDBSCAN istatistiksel olarak daha sağlam ve otomatik seçimli bir yapıya sahiptir.

Neden HDBSCAN Kullanılır?

Özellikle aşağıdaki durumlarda HDBSCAN, DBSCAN'e göre daha iyi sonuçlar verebilir:

- Veri setinde **farklı yoğunluk seviyelerine** sahip kümeler varsa,
- ε parametresini belirlemek zorsa,
- Detaylı bir hiyerarşik kümeleme isteniyorsa,
- Gürültüden etkilenmeyen ve istikrarlı sonuçlar elde edilmek isteniyorsa.