Hierarchical Clustering Algoritması

Atıl Samancıoğlu

1 Giriş

Hierarchical Clustering (Hiyerarşik Kümeleme), veri noktalarını benzerliklerine göre hiyerarşik bir yapı içinde gruplamayı amaçlayan gözetimsiz bir öğrenme algoritmasıdır. K-means gibi sabit bir k değeri baştan verilmek zorunda değildir. Bunun yerine tüm veriler küçük kümelerden başlayarak birleştirilir ya da büyük küme parçalara bölünür.

Bu yöntemin iki temel tipi vardır:

- **Agglomerative Clustering (Alt Alta Büyüme)**: Küçük kümeler birleşerek büyük kümeler oluşturur.
- **Divisive Clustering (Üstten Bölünme)**: Tüm veri bir küme olarak alınır ve adım adım alt kümelere bölünür.

Bu dokümanda ağırlıklı olarak agglomerative yaklaşımı anlatacağız.

2 Temel Adımlar

Hierarchical Clustering üç temel adımdan oluşur:

1. Her veri noktası kendi başına bir kümedir.
2. En yakın iki küme birleştirilir.
3. Tüm veriler tek bir kümede toplanana kadar adım 2 tekrarlanır.

3 Örnek: 7 Nokta Üzerinde Agglomerative Kümeleme

Elimizde P_1'den P_7'ya kadar 7 nokta olsun. Bu noktaların başlangıçta hepsi kendi kümesindedir. İlk olarak birbirine en yakın iki nokta birleştirilir. Sonrasında bu süreç iteratif olarak tekrar edilerek, bir büyük küme oluşana kadar devam eder.

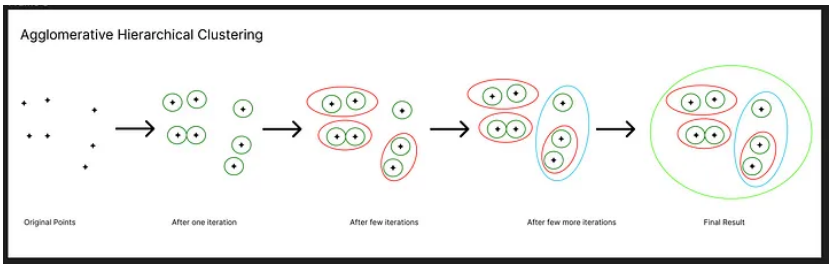

Figure 1: Kaynak: pub.towardsai.net

4 Dendrogram ile Görselleştirme

Yukarıdaki adımlar bir **dendrogram** adlı diyagramla görselleştirilir. Dikey eksen kümeler arası mesafeyi gösterirken yatayda veri noktaları yer alır.

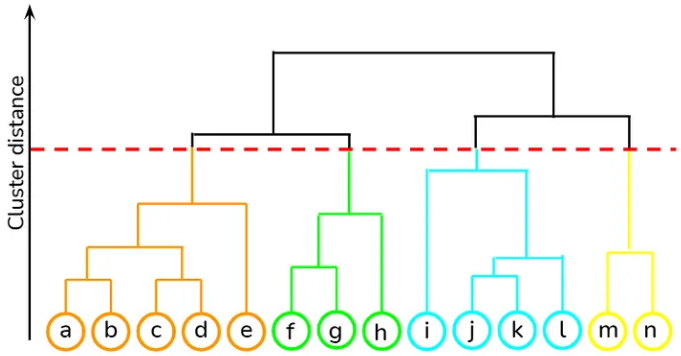

Figure 2: Kaynak: towardsdatascience

Küme Sayısını Belirleme (K Değeri)

Hierarchical Clustering'de küme sayısını belirlemek için en yaygın yöntemlerden biri **dendrogram** kullanmaktır. Bunun için genellikle aşağıdaki adımlar izlenir:

- Dendrogram üzerinde bir **eşik (threshold)** belirlenir. Bu eşik, yatay bir çizgi olarak çizilir.
- Bu çizgi, dendrogramdaki **bağlantıların (merge) seviyelerini** kestiği noktalara göre küme sayısını belirler.
- Yani, bu çizgi kaç tane bağlantıyı kesiyorsa, o kadar küme vardır.

Amaç: Birbirine en uzak (yüksek mesafede birleşen) kümeleri ayırmak.

Bu yöntemin arkasındaki sezgisel fikir şudur: Kümeleme işlemi sırasında bazı kümeler, diğerlerine kıyasla çok daha uzakta birleşir. Bu nedenle bu birleşme noktaları "doğal kümelenme" sınırlarını işaret edebilir.

En yaygın strateji:

Dendrogram'da yatay bir çizgi çekilir ve bu çizgiyle kesilen dikey çizgi sayısı k değerini verir. Bu çizgi:

- En uzun dikey çizgiler arasında,
- Üzerinden en az sayıda yatay birleşmenin geçtiği noktada seçilmelidir.

Alternatif Yöntemler:

- **Gap Statistic**: Her küme sayısı için bir skor hesaplanır ve en büyük "gap" olan yer seçilir.
- **Silhouette Coefficient**: Farklı k değerleri denenerek hangi k'nın küme içi benzerliği en yüksek, küme dışı farkı en yüksek olduğu hesaplanır.
- **Inconsistency Coefficient**: Dendrogram'daki bağlantıların tutarsızlık skorlarına göre küme sınırları belirlenebilir.

Sonuç: Dendrogram üzerindeki kesme çizgisi basit ve sezgisel bir yöntemdir, fakat daha teknik uygulamalarda yukarıdaki yöntemler de tercih edilebilir.

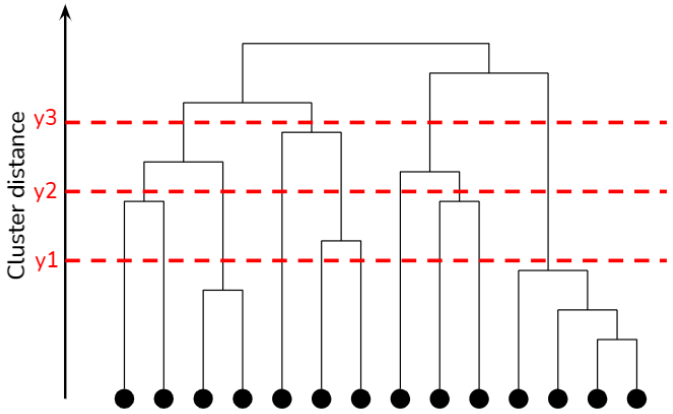

Figure 3: Kaynak: towardsdatascience

5 Mesafe Ölçümleri

Benzerlik hesaplamak için en sık kullanılan ölçütler şunlardır:

Euclidean Distance

İki nokta arası doğrudan uzaklıktır:

$$d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

Manhattan Distance

Sadece yatay ve dikey adımlarla ilerlenir:

$$d = |x_2 - x_1| + |y_2 - y_1|$$

Cosine Similarity

Sayısal olmayan verilerde veya açısal yakınlık hesaplarında kullanılır:

$$\cos(\theta) = \frac{A \cdot B}{||A||\ ||B||}$$

6 Divisive Clustering

Divisive yaklaşımı, tüm veriyi tek bir büyük küme kabul edip, bu kümeyi giderek daha küçük parçalara ayırır. Hesaplama maliyeti daha yüksektir. Bu nedenle çoğunlukla agglomerative tercih edilir.

7 K-Means ve Hierarchical Clustering Karşılaştırması

Özellik	**K-Means**	**Hierarchical Clustering**
Veri Türü	Sadece sayısal	Sayısal + benzerlik hesaplanabilen
Ölçeklenebilirlik	Büyük veri için uygundur	Küçük veri setleri için daha uygundur
Görsellik	Elbow yöntemi ile belirlenir	Dendrogram ile küme sayısı kolayca gözlenebilir
Başlangıç Noktası	Rastgele centroidler	Her nokta kendi kümesidir
Kümeleme Yapısı	Düz yapılı	Hiyerarşik

Table 1: K-Means ve Hierarchical Clustering Karşılaştırması

8 Sonuç

Hierarchical Clustering, özellikle görsel olarak küme yapısını anlamak istediğimiz durumlarda çok güçlüdür. Özellikle:

- Küçük veri setlerinde,
- Karmaşık ilişkili veri noktalarında,
- Sayısal olmayan ama benzerlik tanımlanabilen verilerde,

dendrogram yapısıyla avantaj sağlar.

K-Means, büyük veri setleri için daha uygundur. Doğru algoritma seçimi, problemin veri yapısına ve analiz amacına göre belirlenmelidir.