4.2. Kümeleme (Clustering)

Kümeleme, denetimsiz öğrenmenin bir yöntemidir ve birçok alanda kullanılan istatistiksel veri analizi için yaygın bir tekniktir. Denetimsiz öğrenme, veri kümesi ile çıktıların olmadığı bir öğrenme metodudur. Veri kümesindeki verileri yorumlayarak ortak noktaları bulmak ve bunları kümeleştirme işlemi yapılarak anlamlı bir veri elde edebilmektir. Sistem, öğreten olmadan öğrenmeye çalışır. Ham verileri organize verilere dönüştüren bir makine öğrenimi türüdür.

Denetimsiz öğrenmede sadece veriler vardır onlar hakkında bilgi verilmez. Bu verilerden sonuçlar çıkarılmaya çalışılır. En baştan veriler hakkında herhangi bir bilgi verilmediği için çıkartılan sonuçların kesinlikle doğru olduğu söylenemez. Veriyi değişkenler arasındaki ilişkilere dayalı olarak kümeleyerek çeşitli modeller, yapılar oluşturabiliriz.

Kümelemenin uygulama alanları:

Tıp'da elde edilen görüntülemeler üzerindeki farklıları analiz edilerek değişik nitelikler çıkartılabilir.

Suç Yerlerinin Belirlenmesi: Bir şehirdeki belirli bölgelerde mevcut olan suçlarla ilgili veriler, suç kategorisi, suç alanı ve ikisi arasındaki ilişki, bir şehirdeki ya da bölgedeki suça eğilimli alanlara ilişkin kaliteli bilgiler verebilir.

Oyuncu istatistiklerini analiz etmek: Oyuncu istatistiklerini analiz etmek, spor dünyasının her zaman kritik bir unsuru olmuştur ve artan rekabetle birlikte, makine öğrenmenin burada oynayacağı kritik bir rol vardır.

Kümeleme Çeşitleri:

 Hiyerarşik Kümeleme

 Gürültülü Uygulamaların Yoğunluğa Dayalı Konumsal Kümelenmesi (DBSCAN) (DBSCAN)

 K-means Kümeleme

 Ağırlık Ortalama Kaydırma Kümelemesi

 Gauss Karışım Modelleri (GMM) kullanarak Beklenti-Maksimizasyon (EM) Kümeleme

Hiyerarşik Kümeleme:

Hiyerarşik kümeleme algoritmaları 2 kategoriye ayrılır: yukarıdan aşağıya veya aşağıdan yukarıya. Aşağıdan yukarıya algoritmalar, her veri noktasını başlangıçta tek bir küme olarak ele alır ve ardından tüm kümeler tüm veri noktalarını içeren tek bir kümede birleştirilene kadar küme çiftlerini art arda birleştirir (veya toplar). Bu nedenle aşağıdan yukarıya hiyerarşik kümeleme, hiyerarşik kümelemeli kümeleme (hierarchical agglomerative clustering) veya HAC olarak adlandırılır. Bu küme hiyerarşisi bir ağaç (veya dendrogram) olarak temsil edilir. Ağacın kökü, tüm örnekleri toplayan benzersiz kümedir, yapraklar yalnızca bir örnek içeren kümelerdir.

K-Means Algoritması:

K-means algoritmasında kullanılan örneklem, k adet kümeye bölünür. Algoritmanın özü birbirlerine benzerlik gösteren verilerin aynı küme içerisine alınmasına dayanır. Algoritmadaki benzerlik terimi, veriler arasıdaki uzaklığa göre belirlenmektedir. Uzaklığın az olması benzerliğin yüksek, çok olması ise düşük olduğu anlamına gelmektedir.

K-means algoritmasının yapısı aşağıdaki gibidir;

1) K adet rastgele küme oluştur

2) Kare hata oranını hesapla

3) Verilerin kümelerin orta noktalarına olan uzaklıklarını bul

4) Her veri için en yakın kümeyi, o verinin kümesi olarak belirle

5) Yeni yerleşim düzenine göre hata oranını hesapla

6) Eğer önceki hata oranı ile şimdiki hata oranı eşit değilse 2,3,4,5 ve 6. adımları tekrarla

7) Eğer önceki hata oranı ile şimdiki hata oranı eşitse kümeleme işlemini sonlandır

Dirsek yöntemi, kümelere nasıl ihtiyaç duyacağımız konusunda kullanışlı olur. Dirsek noktasının belirlenmesi gerekir.

Ağırlık Ortalama Kaydırma Kümelemesi:

Ortalama kaydırma kümeleme, veri noktalarının yoğun alanlarını bulmaya çalışan kayan pencere tabanlı bir algoritmadır. Centroid tabanlı bir algoritmadır, yani amacın her bir grubun / sınıfın merkez noktalarını bulmaktır, bu da kayan pencere içindeki noktaların ortalaması olacak merkez noktaları için adayları güncelleyerek çalışır. Bu aday pencereler daha sonra, neredeyse kopyaları ortadan kaldırmak için bir işlem sonrası aşamasında filtrelenir ve nihai merkez noktaları ve bunlara karşılık gelen grupları oluşturur.

Sürgülü pencerelerin tümü ile uçtan uca tüm sürecin bir örneği aşağıda gösterilmiştir. Her siyah nokta, kayan bir pencerenin merkezini temsil eder ve her gri nokta bir veri noktasıdır.

Gürültülü Uygulamaların Yoğunluğa Dayalı Konumsal Kümelenmesi (DBSCAN):

DBSCAN, ortalama kaymaya ekseninde, benzer ve yoğunluklu bölgeleri kümeleyen bir algoritmadır, ancak birkaç önemli avantajı vardır. Minimum bölge sayısında ve uzaklıkta maksimum yoğunluk bölgesi oluşturulması hedeflenir.

Gauss Karışım Modelleri (GMM) kullanarak Beklenti-Maksimizasyon (EM) Kümeleme:

K-Ortalamalarının en büyük dezavantajlarından biri, küme merkezi için ortalama değerin naif kullanımıdır. Aşağıdaki resme bakarak bunun bir şeyleri yapmanın en iyi yolu olmadığını anlayabiliriz. Sol tarafta, aynı ortalamaya merkezlenmiş farklı yarıçaplara sahip iki dairesel küme olduğu insan gözüne oldukça açık görünüyor. K-Ortalamalar bunun üstesinden gelemez çünkü kümelerin ortalama değerleri birbirine çok yakındır. K-Ortalamalar, yine ortalamanın küme merkezi olarak kullanılması sonucunda kümelerin dairesel olmadığı durumlarda başarısız olur.