# Büyük Veri İçin İstatistiksel Öğrenme (Statistical Learning for Big Data)

M. Ali Akcayol
Gazi Üniversitesi
Bilgisayar Mühendisliği Bölümü

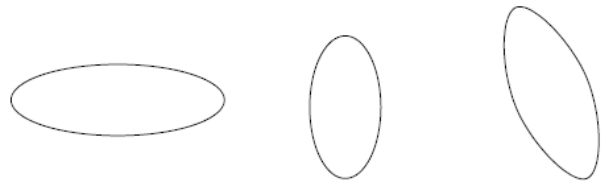
Bu dersin sunumları, "The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Trevor Hastie, Robert Tibshirani, Jerome Friedman, Springer, 2017." ve "Mining of Massive Datasets, Jure Leskovec, Anand Rajaraman, Jeffrey David Ullman, Stanford University, 2011." kitapları kullanılarak hazırlanmıştır.

## Konular

- BFR algoritması
  - BFR algoritmasının çalışması
- CURE Algoritması
  - CURE algoritmasında başlangıç
  - CURE algoritmasının tamamlanması

## BFR algoritması

- **BFR** (Bradley, Fayyad, Reina), k-means algoritmasının **çok boyutlu Öklit uzayında clustering** için tasarlanmış şeklidir.
- BFR algoritması cluster içindeki **noktaların centroid noktasına göre düzgün dağılımda** olduğunu kabul eder.
- Cluster içindeki noktaların boyutlara göre ortalaması ve standart sapması farklı olabilir, ancak **cluster eksenleri Öklit uzayındaki eksenlerle aynı olmalıdır.**



OK OK Not OK

3

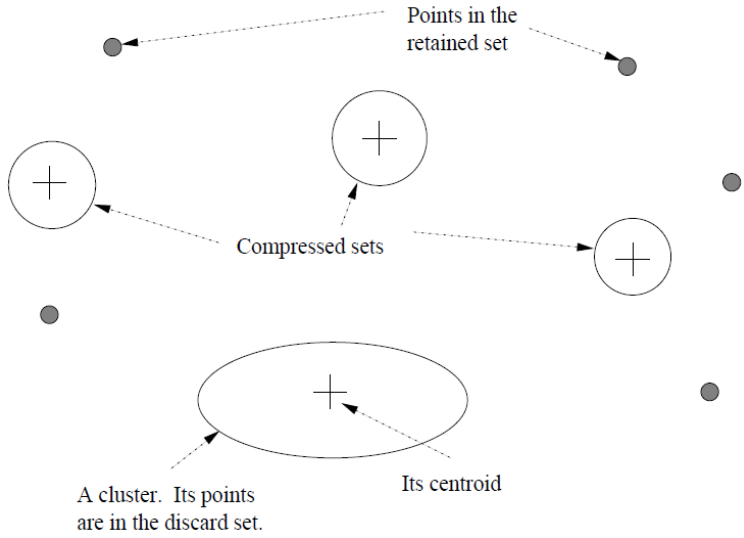
## BFR algoritması

- BFR algoritması ***k* tane noktayı seçerek başlar.**
- Başlangıç noktaları olabildiği kadar birbirinden uzak seçilir.
- Veri dosyasından **noktalar büyük bloklar halinde okunur.**
- Her veri bloğu **hafızada işlenebilecek kadar noktaya sahiptir.**
- **Hafızada *k* tane cluster özeti saklanır.**
- Ana hafızada veri bloğu dışında **üç tür nesne vardır:**
  - **Discard set:** Cluster'ların **basit özetleridir.** Özetin gösterdiği noktalar atılır, ancak özet hafızada tutulur.
  - **Compressed set:** Cluster özetine benzer şekilde **birbirine yakın noktaların özetidir.** Diğer cluster'lara yakın değillerdir. Bu cluster özetine minicluster denir.
  - **Retained set:** Herhangi bir **cluster'a** henüz **atanmamış noktalar kümesidir.** Herhangi bir compressed set içerisine alınacak kadar yakın değillerdir. Bu noktalar hafızada olduğu gibi tutulur.

4

## BFR algoritması

- BFR algoritmasındaki **discard**, **compressed** ve **retained set** gösterimi.




## BFR algoritması

- Cluster'lar için **discard** ve **compressed kümeleri** $d$ boyutlu uzayda $\mathbf{2d + 1}$ **tane değer ile gösterilir.**
  - **Temsil edilen noktaların sayısı** $N$ ile gösterilir.
  - Her boyuttaki **tüm noktaların bileşenlerinin toplamı.** $d$ **boyutunda bir vektördür** (SUM). ***i.*boyuttaki toplam değer** $\mathbf{SUM}_i$ **olur.**
  - Her boyuttaki **tüm noktaların bileşenlerinin karelerinin toplamı.** $d$ **boyutunda bir vektördür** (SUMSQ). ***i.*boyuttaki toplam değer** $\mathbf{SUMSQ}_i$ **olur.**
- Bir nokta kümesi her boyut için, **nokta adedi**, **centroid** ve **standart sapma** ile gösterilir.
- $i$.boyuttaki **centroid koordinatı** $\mathrm{SUM}_i / N$ ile hesaplanır.
- $i$.boyuttaki **varyans** $(\mathrm{SUMSQ}_i / N) - (\mathrm{SUM}_i / N)^2$ ile hesaplanır.
- **Standart sapma** ise **varyansın karekökü** alınarak hesaplanır.

## BFR algoritması

**Örnek**

- Bir cluster (5, 1), (6, -2) ve (7, 0) noktalarına sahip olsun.
- $N$ = 3, **SUM** = [18, -1], **SUMSQ** = [110, 5] olur.
- Centroid = SUM / $N$ = [6, -1/3] olur.
- **Birinci boyut için varyans** = $110/3 - (18/3)^2$ = **0.667**, **Standart sapma = 0.816** olur.
- **İkinci boyut için varyans** = $5/3 - (-1/3)^2$ = **1.56, Standart sapma = 1.25** olur.

7

## Konular

- BFR algoritması
  - BFR algoritmasının çalışması
- CURE Algoritması
  - CURE algoritmasında başlangıç
  - CURE algoritmasının tamamlanması

8

## BFR algoritmasının çalışması

- BFR algoritması her yeni gelen **veri bloğu (chunk) içindeki noktalar üzerinde işlem yapar.**
- Bir **cluster centroid'ine** yeterli düzeyde **yakın olan tüm noktalar o cluster'a atanır.**
- Cluster'ın $N$, $\mathbf{SUM}$, $\mathbf{SUMSQ}$ **değerleri güncellenir** ve yeni nokta silinir.
- Yeterli düzeyde bir **cluster'a yakın olmayan** noktalar **retained set içerisine aktarılır.**
- Singleton cluster'lar retained set noktalarını oluşturur.
- Retained set içerisindeki noktalar hiyerarşik olarak kümelenebilir.
- **Birbirine yeterli düzeyde yakın iki nokta bir cluster'a dönüştürülür.**
- **Birbirine yeterli düzeyde yakın iki cluster bir cluster'a dönüştürülür.**

9

## BFR algoritmasının çalışması

- **Birden fazla noktaya sahip cluster'lar** özetlenir ve **compressed set oluşturulur.**
- Singleton cluster'lar retained set içerisindeki noktaları oluşturur.
- Bir **cluster'a atanan noktalar ikincil diske aktarılır** ve hafızadan silinir.
- Minicluster'lar ve retained set noktaları $k$ tane cluster'dan birisine belirli düzeyde yakın değilse birleştirilemez.

10

## BFR algoritmasının çalışması

- **Son chunk geldikten sonra**, **compressed set** ve **retained set üzerinde** bazı işlemlerin yapılması gerekir:
  1. Retained set noktaları **outlier olarak alınabilir** ve **bir cluster'a atanmaz.**
  2. Retained set noktaları **kendisine en yakın cluster'a atanabilir.**
- Yeni gelen noktanın hangi cluster'a atanacağı iki şekilde belirlenebilir:
  - Bir $p$ noktası **kendisine en yakın centroid'e sahip cluster'a atanır.** Ancak, chunk içindeki **tüm noktalar atandığında başka bir cluster centroid'i $p$ noktasına daha yakın olabilir.** Tüm noktalardan örnekler seçilerek tahmin yapılır.
  - $p$ noktasının cluster'ların centroid'i arasındaki Mahalanobis uzaklığı hesaplanarak ait olacağı cluster tahmin edilir. $p = [p_1, p_2, \ldots, p_d]$, $c = [c_1, c_2, \ldots, c_d]$, $\sigma$ = stddev.

$$\text{Mahalanobis distance} = \sqrt{\sum_{i=1}^{d} \left(\frac{p_i - c_i}{\sigma_i}\right)^2}$$

11

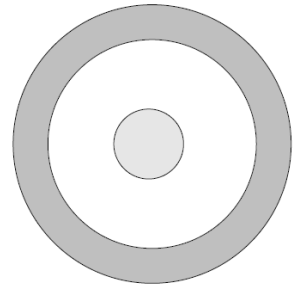
## Konular

- BFR algoritması
  - BFR algoritmasının çalışması
- CURE Algoritması
  - CURE algoritmasında başlangıç
  - CURE algoritmasının tamamlanması

12

## CURE algoritması

- **CURE (Clustering Using REpsentatives) algoritması** Öklit uzayında **büyük ölçekli veride clustering yapar.**
- CURE algoritması **az sayıda örnek noktayı** kullanarak **diskteki büyük veriyi cluster'lara ayırır.**
- Cluster'ları centroid ile temsil etmek yerine, **bir grup nokta ile temsil eder.** Şekilde iki cluster ile gösterim yapılmıştır.



- İlk cluster daire, ikinci cluster etrafını saran çember ile ifade edilmiştir.

13

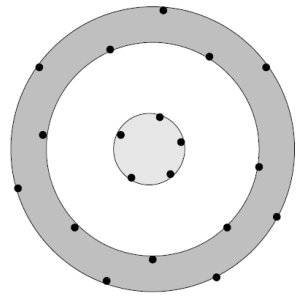
## Konular

- BFR algoritması
  - BFR algoritmasının çalışması
- CURE Algoritması
  - CURE algoritmasında başlangıç
  - CURE algoritmasının tamamlanması

14

## CURE algoritmasında başlangıç

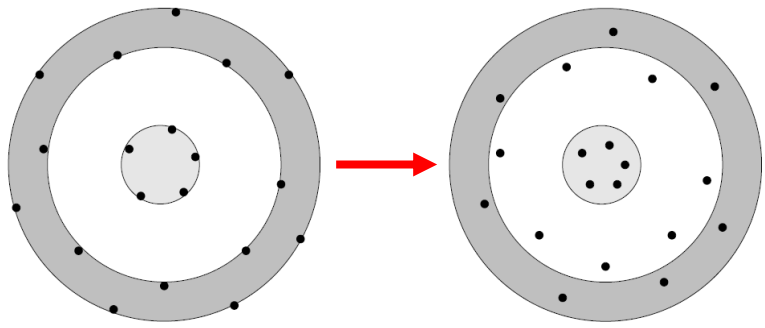
- **Küçük bir örnek veri hafızaya alınır** ve **cluster'lar oluşturulur.**
- Genellikle orijinal verinin %2.5 kısmı rastgele seçilerek alınabilmektedir.
- **Başlangıçta herhangi bir clustering yöntemi kullanılabilir.** CURE algoritması tek sayıda cluster oluşturur (Hiyerarşik gösterim için).
- **Her cluster az sayıda nokta kümesi seçilerek temsil edilir.**
- Seçilen **noktalar** olabildiği kadar **birbirinden uzakta olmalıdır.**



15

## CURE algoritmasında başlangıç

- Cluster'ı **temsil eden noktalar**, centroid ile aralarındaki uzaklığın belirli oranında **centroid noktasına doğru yer değiştirir** ($0 \leq \alpha \leq 1$).
- Temsil noktaları $\alpha \leq 0.2$ (%20) oranında yer değiştirilebilir.



- **Farklı iki cluster'daki en yakın temsil noktaları** birbirine **eşik değerden daha yakın olan cluster'lar birleştirilir.**

16

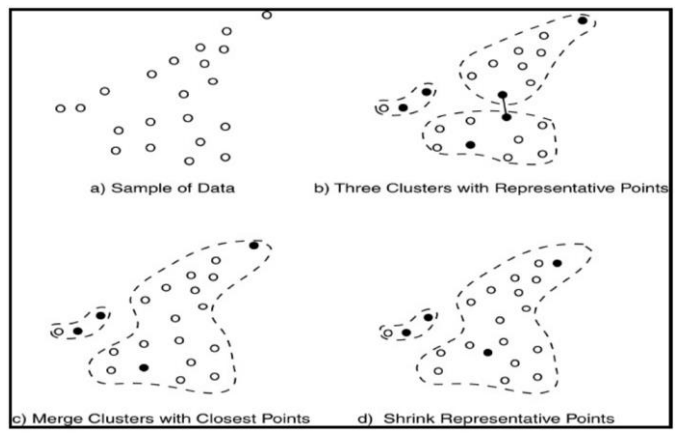
## Konular

- BFR algoritması
  - BFR algoritmasının çalışması
- CURE Algoritması
  - CURE algoritmasında başlangıç
  - CURE algoritmasının tamamlanması

17

## CURE algoritmasının tamamlanması

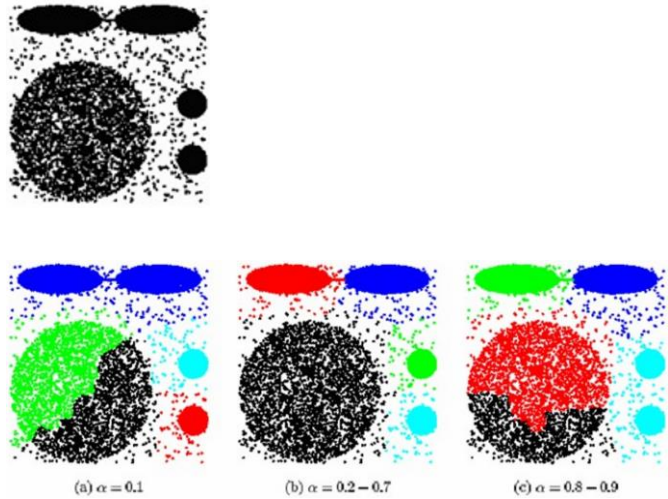
- Noktaların yer değiştirmesinden sonra birbirine yeterince yaklaşan iki ayrı nokta (farklı iki cluster'da) varsa bu cluster'lar birleştirilir.
- **Birleştirme** aşaması **yeterince yakın iki cluster bulunamayıncaya kadar tekrarlanır.**




18

## CURE algoritmasının tamamlanması

**Örnek**




19

## Ödev

- Stream'ler için clustering algoritmaları hakkında bir araştırma ödevi hazırlayınız.

20