Büyük Veri Analitiği
(Big Data Analytics)

M. Ali Akcayol
Gazi Üniversitesi
Bilgisayar Mühendisliği Bölümü

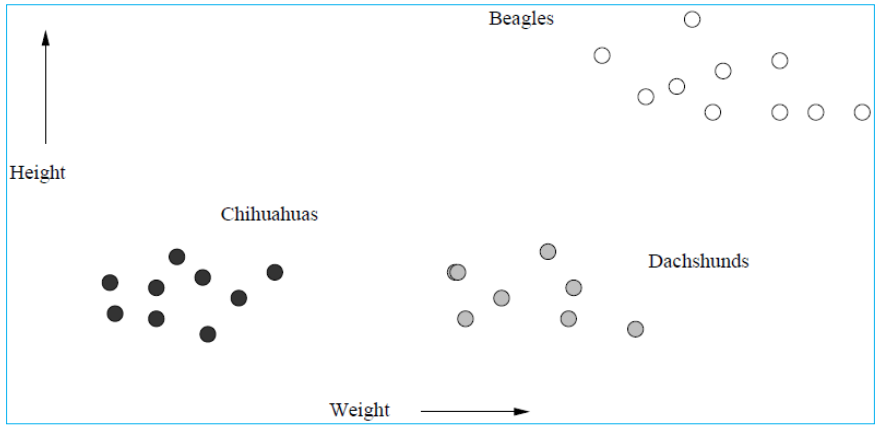
Bu dersin sunumları, "Mining of Massive Datasets, Jure Leskovec, Anand Rajaraman, Jeffrey David Ullman, Stanford University, 2011." kitabı kullanılarak hazırlanmıştır.

Konular

- Clustering Yöntemleri
  - Points, spaces, distances
  - Clustering stratejileri
- Hiyerarşik Clustering
  - Öklit uzayında hiyerarşik clustering
  - Öklit olmayan uzaylarda hiyerarşik clustering
- K-Means Algoritması
  - K-means için başlangıç cluster'ları
  - Doğru k değerinin belirlenmesi
- BFR Algoritması
  - BFR algoritmasının çalışması
- CURE Algoritması
  - CURE algoritmasında başlangıç
  - CURE algoritmasının tamamlanması

Clustering yöntemleri

- **Clustering noktalar topluluğunun** bir uzaklık ölçütüne göre **gruplandırılmasıdır.**
- Aynı cluster içerisinde yer alan noktalar diğer cluster'lar içerisinde yer alan noktalara göre birbirine daha yakındır.



3

Konular

- Clustering Yöntemleri
  - Points, spaces, distances
  - Clustering stratejileri
- Hiyerarşik Clustering
  - Öklit uzayında hiyerarşik clustering
  - Öklit olmayan uzaylarda hiyerarşik clustering
- K-Means Algoritması
  - K-means için başlangıç cluster'ları
  - Doğru k değerinin belirlenmesi
- BFR Algoritması
  - BFR algoritmasının çalışması
- CURE Algoritması
  - CURE algoritmasında başlangıç
  - CURE algoritmasının tamamlanması

4

Points, spaces, distances

- Clustering için iki temel yaklaşım vardır: **hiyerarşik** ve **nokta atama**.
- Clustering için uygun **bir veri seti noktalar topluluğudur** ve **her nokta uzaydaki bir nesnedir.**
- **Öklit uzayındaki noktalar reel sayılardan oluşan vektör ile gösterilir.**
- Vektör elemanları koordinat olarak adlandırılır.
- Günümüzdeki **clustering problemleri** çok **büyük boyuttadır.**
- Noktalar arasındaki uzaklık ölçütlerinde aşağıdaki şartlar sağlanır:
  - **Noktalar arasındaki uzaklıklar her zaman pozitif olur.**
  - **Uzaklık simetriktir.** Uzaklık hesaplanırken noktaların sırası önemli değildir.
  - **Uzaklık ölçütleri üçgen eşitsizliğine uymalıdır.** $d(x, y) + d(y, z) \geq d(x, z)$

5

Konular

- Clustering Yöntemleri
  - Points, spaces, distances
  - Clustering stratejileri
- Hiyerarşik Clustering
  - Öklit uzayında hiyerarşik clustering
  - Öklit olmayan uzaylarda hiyerarşik clustering
- K-Means Algoritması
  - K-means için başlangıç cluster'ları
  - Doğru k değerinin belirlenmesi
- BFR Algoritması
  - BFR algoritmasının çalışması
- CURE Algoritması
  - CURE algoritmasında başlangıç
  - CURE algoritmasının tamamlanması

6

Clustering stratejileri

- **Cluster özeti** için **Öklit uzayında** noktaların orta noktası (centroid) alınır.
- **Öklit dışındaki uzaylarda cluster özeti için farklı yöntemler kullanılır.**
- Kullanılan yöntemlere göre **clustering algoritmaları iki gruba ayrılır:**

(1) Hiyerarşik veya agglomerative

- **Her nokta kendi cluster'ına ait tek nokta** alınarak olarak başlanır.
- Yakınlık durumuna göre **noktalar birleştirilerek cluster'lar oluşturulur.**
- Algoritma **önceden belirlenen cluster sayısına ulaşıldığında** veya **noktalar arasında belirli uzaklığa ulaşıldığında sonlanır.**

(2) Nokta atama

- Başlangıçta **belirlenen sayıda cluster belirlenir.**
- **Her nokta en iyi eşleştirildiği cluster'a atanır.**
- Outlier noktalar herhangi bir cluster'a atanmayabilir.

7

Konular

- Clustering Yöntemleri
  - Points, spaces, distances
  - Clustering stratejileri
- Hiyerarşik Clustering
  - Öklit uzayında hiyerarşik clustering
  - Öklit olmayan uzaylarda hiyerarşik clustering
- K-Means Algoritması
  - K-means için başlangıç cluster'ları
  - Doğru k değerinin belirlenmesi
- BFR Algoritması
  - BFR algoritmasının çalışması
- CURE Algoritması
  - CURE algoritmasında başlangıç
  - CURE algoritmasının tamamlanması

8

Hiyerarşik Clustering

- Hiyerarşik clustering algoritmalarında **her nokta bir cluster alınarak başlanır** ve **cluster'lar birleştirilir.**
- Öklit uzayında **cluster'ların özetleri için centroid kullanılır.**
- **Öklit olmayan uzaylarda** ise **cluster'ların özeti için** clustroid **kullanılır.**
- **Clustroid bir cluster'ı temsil eder** ve uygulamaya özgü belirlenecek bir yöntemle elde edilebilir.

9

Konular

- Clustering Yöntemleri
  - Points, spaces, distances
  - Clustering stratejileri
- Hiyerarşik Clustering
  - Öklit uzayında hiyerarşik clustering
  - Öklit olmayan uzaylarda hiyerarşik clustering
- K-Means Algoritması
  - K-means için başlangıç cluster'ları
  - Doğru k değerinin belirlenmesi
- BFR Algoritması
  - BFR algoritmasının çalışması
- CURE Algoritması
  - CURE algoritmasında başlangıç
  - CURE algoritmasının tamamlanması

10

Öklit uzayında hiyerarşik clustering

- Tüm **hiyerarşik clustering algoritmaları her noktayı bir cluster alarak başlar.**
- Küçük **iki cluster birleştirilerek büyük bir cluster oluşturulur.**
- Hiyerarşik clustering algoritmalarında aşağıdakilerin belirlenmesi gerekir:
  - **Cluster'ların nasıl gösterileceği**
  - **İki cluster'ın birleştirilmesinin nasıl yapılacağı**
  - **Cluster birleştirmenin ne zaman sonlanacağı**
- Algoritma aşağıdaki işlem adımlarını tekrarlar.

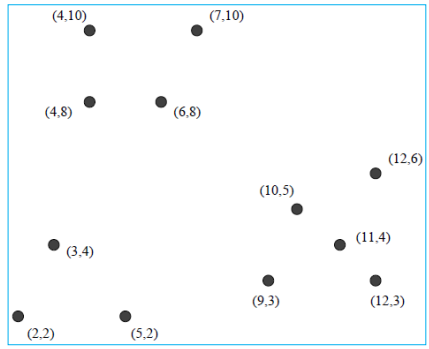
```
WHILE it is not time to stop DO
    pick the best two clusters to merge;
    combine those two clusters into one cluster;
END;
```

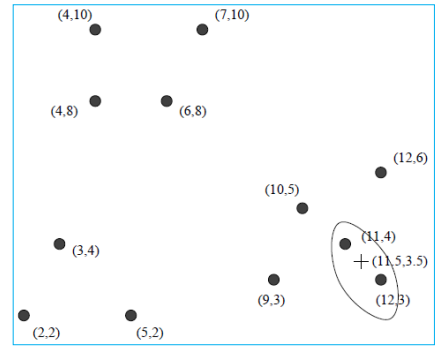
11

Öklit uzayında hiyerarşik clustering

Örnek

- Aşağıdaki veri seti iki boyutlu Öklit uzayındadır.
- **Başlangıçta tüm noktalar** kendi cluster'ına aittir ve **orta noktadır.**
- En yakın iki nokta çifti **(10, 5) ile (11, 4)** ve **(11, 4) ile (12, 3),** $d = 2^{1/2}$
- (11, 4) ile (12, 3) birleştirildiğinde orta nokta (11.5, 3.5) olur.



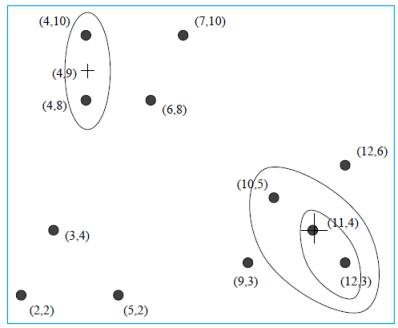


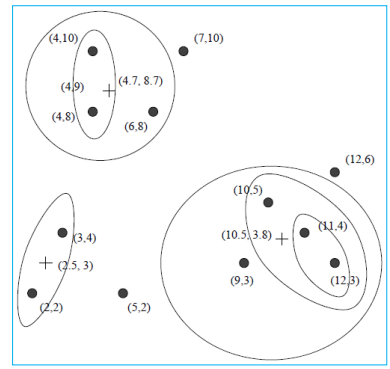
12

Öklit uzayında hiyerarşik clustering

Örnek

- Sonraki adımda **(10, 5) ile centroid arasındaki uzaklık (2.12) olur.**
- **(4, 8) ile (4, 10) arasındaki uzaklık (2.0) olur.** İkinci adımda bu iki nokta birleştirilir ve **centroid (4, 9).**
- Ardından **(10, 5) noktası ile (11.5, 3.5) cluster'ı birleştirilir.**





13

Öklit uzayında hiyerarşik clustering

Cluster birleştirme kuralları

- İki farklı cluster içindeki noktalardan **en yakın olanların uzaklığı minimum olan iki cluster birleştirilir** (En yakın komşu algoritması).
- İki farklı cluster içindeki noktalardan **en uzak olanların uzaklığı minimum olan iki cluster birleştirilir** (En uzak komşu algoritması).
- İki farklı cluster'daki **tüm nokta çiftlerinin birbirine uzaklıklarının ortalaması minimum olan iki cluster birleştirilir.**
- **Bir cluster'ın yarıçapı tüm noktaların centroid'e maksimum uzaklığını belirler.** İki cluster birleştirilirken **minimum yarıçapı oluşturacak iki cluster birleştirilir.**
- **Bir cluster'ın çapı cluster içindeki en uzak iki noktanın uzaklığını belirler.** İki cluster birleştirilirken **minimum çapı oluşturacak iki cluster birleştirilir.**

14

Öklit uzayında hiyerarşik clustering

Cluster birleştirmeyi sonlandırma

- **Cluster çapı** belirlenen **threshold değerini aştığında** birleştirme yapılmaz.
- **Cluster içindeki nokta yoğunluğu** belirlenen **threshold değerini aştığında** birleştirme yapılmaz.
- İki cluster birleştirilince **kötü bir cluster oluşacaksa** (Örn.: Cluster çapı aniden çok yükselecekse) **birleştirme yapılmaz.**

15

Öklit uzayında hiyerarşik clustering

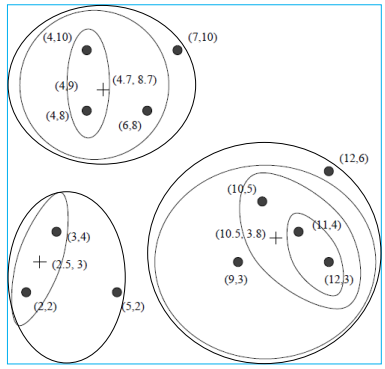
Algoritmayı sonlandırma

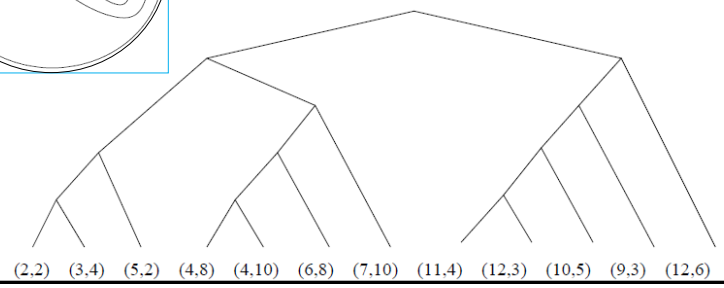
- Önceden **belirlenen sayıda cluster'a ulaşıldığında algoritma sonlandırılabilir.**
- Cluster centroid **noktasından ortalama uzaklık belirli bir threshold değeri aştığında sonlandırılabilir.** Cluster genişliği belirli bir alanda tutulmak istenebilir.
- **Tüm cluster'lar birleştirilip tek cluster elde edilince** algoritma sonlandırılır.

16

Öklit uzayında hiyerarşik clustering

- Tüm cluster'lar ağaç yapısında gösterilebilir.





17

Konular

- Clustering Yöntemleri
  - Points, spaces, distances
  - Clustering stratejileri
- Hiyerarşik Clustering
  - Öklit uzayında hiyerarşik clustering
  - Öklit olmayan uzaylarda hiyerarşik clustering
- K-Means Algoritması
  - K-means için başlangıç cluster'ları
  - Doğru k değerinin belirlenmesi
- BFR Algoritması
  - BFR algoritmasının çalışması
- CURE Algoritması
  - CURE algoritmasında başlangıç
  - CURE algoritmasının tamamlanması

18

Öklit olmayan uzaylarda hiyerarşik clustering

- Öklit olmayan uzaylarda **uzaklık ölçütünün** (Jaccard, edit, …) **belirlenmesi gereklidir.**
- Öklit olmayan uzaylarda **konum bilgisi tanımlanmaz.**
- Öklit olmayan uzaylarda **iki noktanın orta noktası da tanımlanamayabilir.**
- Öklit olmayan uzaylarda **noktalar birleştirilemez** ve **noktalardan birisi cluster'ı temsil eder (clustroid).**
- **Clustroid noktası için,**
  - Cluster içindeki **diğer noktalara uzaklığın toplamı alınabilir.**
  - Cluster içindeki **diğer noktalara maksimum uzaklığı olan nokta alınabilir.**
  - Cluster içindeki **diğer noktalara uzaklıkların karelerinin toplamı alınabilir.**
  - **Tüm noktalara en yakın nokta orta nokta seçillir.**

19

Öklit olmayan uzaylarda hiyerarşik clustering

Örnek

- Öklit olmayan uzaylarda uzaklık ölçütünün belirlenmesi gereklidir.
- Bir cluster **abcd**, **aecdb**, **abecb**, **ecdab** noktalarına sahip olsun.
- Aralarındaki uzaklık **edit distance** ile aşağıdaki gibi hesaplanır.

	ecdab	abecb	aecdb
abcd	5	3	3
aecdb	2	2	
abecb	4		

- Cluster'ın clustroid noktası için üç kriter aşağıdaki gibi hesaplanır.

Point	Sum	Max	Sum-Sq
abcd	11	5	43
aecdb	7	3	17
abecb	9	4	29
ecdab	11	5	45

- Üç kritere göre de **aecdb** noktası clustroid alınır.

20

Öklit olmayan uzaylarda hiyerarşik clustering

Cluster birleştirme

- Clustroid noktaları **birbirine en yakın olan cluster'lar birleştirilebilir.**
- Cluster'lardaki **tüm noktaların arasındaki uzaklıkların minimum olduğu iki cluster birleştirilebilir.**
- Cluster'lardaki **noktaların uzaklıklarının ortalamalarının minimum olduğu iki cluster birleştirilebilir.**

Birleştirmenin sonlandırılması

- Cluster içerisindeki **nokta yoğunluğuna göre** birleştirme sonlandırılabilir.
- Cluster **yarıçapı** veya **çapına göre** birleştirme sonlandırılabilir.

21

Konular

- Clustering Yöntemleri
  - Points, spaces, distances
  - Clustering stratejileri
- Hiyerarşik Clustering
  - Öklit uzayında hiyerarşik clustering
  - Öklit olmayan uzaylarda hiyerarşik clustering
- K-Means Algoritması
  - K-means için başlangıç cluster'ları
  - Doğru k değerinin belirlenmesi
- BFR Algoritması
  - BFR algoritmasının çalışması
- CURE Algoritması
  - CURE algoritmasında başlangıç
  - CURE algoritmasının tamamlanması

22

K-Means Algoritması

- **K-means algoritması nokta ataması şeklinde clustering** yapan algoritmalardan en yaygın kullanılandır.
- K-means algoritması **Öklit uzayında clustering yapar** ve **cluster sayısı (k) başlangıçta belirlenir.**

```
Initially choose k points that are likely to be in
    different clusters;
Make these points the centroids of their clusters;
FOR each remaining point p DO
    find the centroid to which p is closest;
    Add p to the cluster of that centroid;
    Adjust the centroid of that cluster to account for p;
END;
```

- Algoritmanın temeli for-loop kısmıdır. **Noktaların en yakın olduğu centroid bulunarak cluster ataması yapılır.**
- Cluster'ın **centroid noktası nokta eklendikçe yeniden hesaplanır.**

23

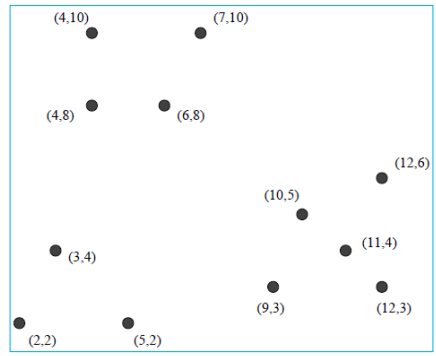
Konular

- Clustering Yöntemleri
  - Points, spaces, distances
  - Clustering stratejileri
- Hiyerarşik Clustering
  - Öklit uzayında hiyerarşik clustering
  - Öklit olmayan uzaylarda hiyerarşik clustering
- K-Means Algoritması
  - K-means için başlangıç cluster'ları
  - Doğru k değerinin belirlenmesi
- BFR Algoritması
  - BFR algoritmasının çalışması
- CURE Algoritması
  - CURE algoritmasında başlangıç
  - CURE algoritmasının tamamlanması

24

K-means için başlangıç cluster'ları

- **Başlangıç noktaları** birbirlerinden **olabildiğince uzak olmalıdır.**
- **Örnek veri üzerinde** hiyerarşik clustering yapılarak centroid noktaları belirlenebilir.
- Örnekte **ilk nokta (6, 8), ikinci nokta (12, 3)** alınabilir.
- İlk iki noktaya en uzak nokta **(2, 2) üçüncü nokta** alınabilir.



25

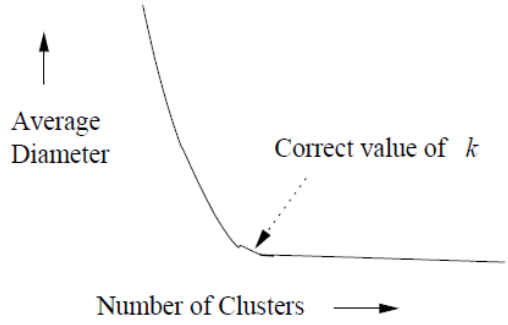
Konular

- Clustering Yöntemleri
  - Points, spaces, distances
  - Clustering stratejileri
- Hiyerarşik Clustering
  - Öklit uzayında hiyerarşik clustering
  - Öklit olmayan uzaylarda hiyerarşik clustering
- K-Means Algoritması
  - K-means için başlangıç cluster'ları
  - Doğru k değerinin belirlenmesi
- BFR Algoritması
  - BFR algoritmasının çalışması
- CURE Algoritması
  - CURE algoritmasında başlangıç
  - CURE algoritmasının tamamlanması

26

Doğru k değerinin belirlenmesi

- Doğru *k* değeri bilinemeyebilir, ancak **farklı *k* değerleri için clustering kalitesi ölçülebilir.**
- Seçilen **cluster sayısı, doğru** cluster sayısına **eşit** veya **büyükse cluster yarıçapı** veya **çapı nokta ekledikçe yavaş bir şekilde artar.**
- Seçilen cluster sayısı, doğru cluster sayısından **küçük** ise **yarıçap** veya **çap aniden yükselir.**



27

Doğru k değerinin belirlenmesi

- Doğru *k* değerine ilişkin bir bilgi yoksa, ***k* değeri** 1, 2, 4, 8, … şeklinde **artırılarak denenir** ve **en uygun *k* değeri belirlenir.**
- **Yarıçap** veya **çap** değeri **hangi aralıkta aniden düşerse** o aralıkta **binary search ile doğru *k* değeri belirlenebilir.**
- *k* değerinin *x* ile *y* arasında olacağı belirlenmiş olsun.
- $z = (x + y) / 2$ değerine bakılır.
- *x* ile *z* arasındaki değişim ile *z* ile *y* arasındaki değişime bakılır.
- Hangi aralıkta değişim yüksekse o aralıkla devam edilir.

28

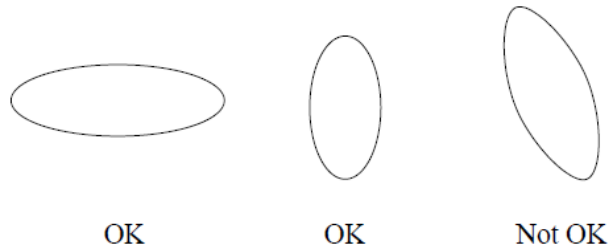
Konular

- Clustering Yöntemleri
  - Points, spaces, distances
  - Clustering stratejileri
- Hiyerarşik Clustering
  - Öklit uzayında hiyerarşik clustering
  - Öklit olmayan uzaylarda hiyerarşik clustering
- K-Means Algoritması
  - K-means için başlangıç cluster'ları
  - Doğru k değerinin belirlenmesi
- BFR Algoritması
  - BFR algoritmasının çalışması
- CURE Algoritması
  - CURE algoritmasında başlangıç
  - CURE algoritmasının tamamlanması

29

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.**



30

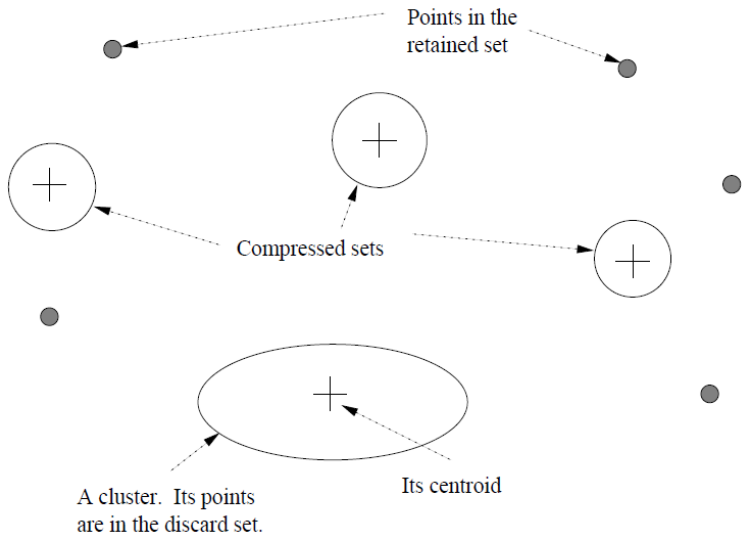
BFR algoritması

- 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 tutulur.

31

BFR algoritması

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



32

BFR algoritması

- Cluster'lar için **discard** ve **compressed kümeleri** d boyutlu uzayda $2d + 1$ **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** 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.

33

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.

34

Konular

- Clustering Yöntemleri
  - Points, spaces, distances
  - Clustering stratejileri
- Hiyerarşik Clustering
  - Öklit uzayında hiyerarşik clustering
  - Öklit olmayan uzaylarda hiyerarşik clustering
- K-Means Algoritması
  - K-means için başlangıç cluster'ları
  - Doğru k değerinin belirlenmesi
- BFR Algoritması
  - BFR algoritmasının çalışması
- CURE Algoritması
  - CURE algoritmasında başlangıç
  - CURE algoritmasının tamamlanması

35

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, **SUM**, **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.**

36

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.

37

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. $p = [p_1, p_2, \ldots, p_d]$, $c = [c_1, c_2, \ldots, c_d]$, σ = stddev.

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

38

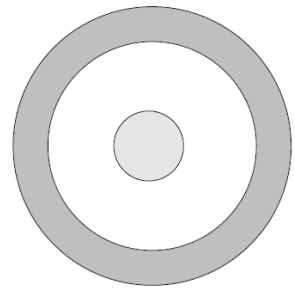
Konular

- Clustering Yöntemleri
  - Points, spaces, distances
  - Clustering stratejileri
- Hiyerarşik Clustering
  - Öklit uzayında hiyerarşik clustering
  - Öklit olmayan uzaylarda hiyerarşik clustering
- K-Means Algoritması
  - K-means için başlangıç cluster'ları
  - Doğru k değerinin belirlenmesi
- BFR Algoritması
  - BFR algoritmasının çalışması
- CURE Algoritması
  - CURE algoritmasında başlangıç
  - CURE algoritmasının tamamlanması

39

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.

40

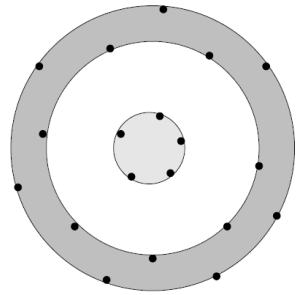
Konular

- Clustering Yöntemleri
  - Points, spaces, distances
  - Clustering stratejileri
- Hiyerarşik Clustering
  - Öklit uzayında hiyerarşik clustering
  - Öklit olmayan uzaylarda hiyerarşik clustering
- K-Means Algoritması
  - K-means için başlangıç cluster'ları
  - Doğru k değerinin belirlenmesi
- BFR Algoritması
  - BFR algoritmasının çalışması
- CURE Algoritması
  - CURE algoritmasında başlangıç
  - CURE algoritmasının tamamlanması

41

CURE algoritmasında başlangıç

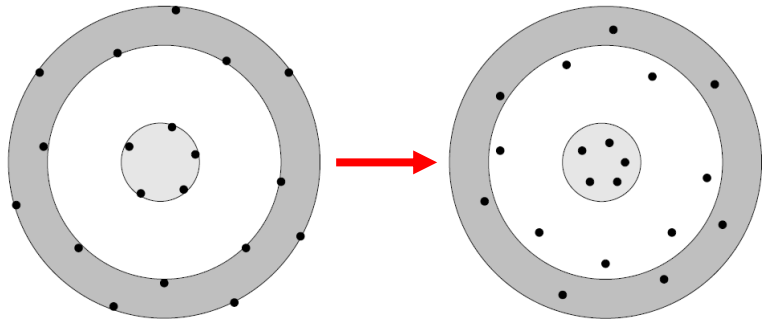
- **Küçük bir örnek veri hafızaya alınır** ve **cluster'lar oluşturulur.**
- Genellikle orijinal verinin %2.5 kısmı rastgele seçilerek alınabilmektedir.
- **Başlangıçta herhangi bir clustering yöntemi kullanılabilir.** CURE algoritması tek sayıda cluster oluşturur (Hiyerarşik gösterim için).
- **Her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



42

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.**

43

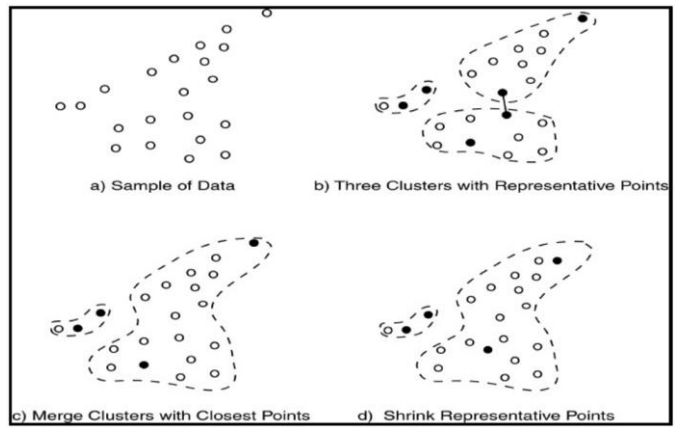
Konular

- Clustering Yöntemleri
  - Points, spaces, distances
  - Clustering stratejileri
- Hiyerarşik Clustering
  - Öklit uzayında hiyerarşik clustering
  - Öklit olmayan uzaylarda hiyerarşik clustering
- K-Means Algoritması
  - K-means için başlangıç cluster'ları
  - Doğru k değerinin belirlenmesi
- BFR Algoritması
  - BFR algoritmasının çalışması
- CURE Algoritması
  - CURE algoritmasında başlangıç
  - CURE algoritmasının tamamlanması

44

CURE algoritmasının tamamlanması

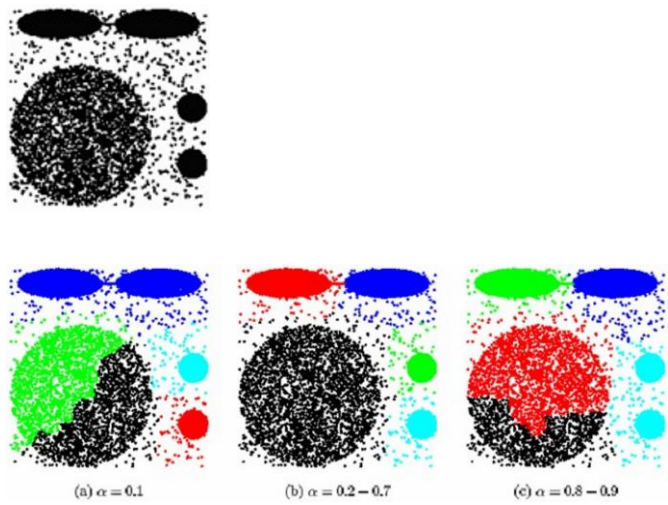
- Noktaların yer değiştirmesinden sonra birbirine yeterince yaklaşan iki ayrı nokta (farklı iki cluster'da) varsa bu cluster'lar birleştirilir.
- **Birleştirme** aşaması **yeterince yakın iki cluster bulunamayıncaya kadar tekrarlanır.**



45

CURE algoritmasının tamamlanması

Örnek



46

Ödev

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

47