

# VERİ MADENCİLİĞİ

## Demetleme Yöntemleri

Yrd. Doç. Dr. Şule Gündüz Öğüdücü  
<http://www3.itu.edu.tr/~sgunduz/courses/verimaden/>

1

## Konular

- Demetleme işlemleri
  - Demetleme tanımı
  - Demetleme uygulamaları
- Demetleme Yöntemleri
  - Bölünmeli Yöntemler
  - Hiyerarşik Yöntemler
  - Yoğunluk Tabanlı Yöntemler
  - Model Tabanlı Yöntemler

<http://www3.itu.edu.tr/~sgunduz/courses/verimaden/>

2

## Demetleme

- Nesneleri demetlere (gruplara) ayırma
- Demet: birbirine benzeyen nesnelerden oluşan grup
  - Aynı demetdeki nesneler birbirine daha çok benzer
  - Farklı demetlerdeki nesneler birbirine daha az benzer

Aynı demet içindeki nesneler arasındaki uzaklığı en küçültme

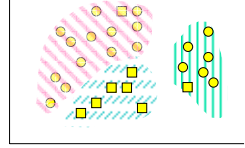
Farklı demetlerdeki nesneler arasındaki uzaklığı en büyütme

<http://www3.itu.edu.tr/~sgunduz/courses/verimaden/>

3

## Demetleme

- Gözetimsiz öğrenme: Hangi nesnenin hangi sınıfa ait olduğu ve sınıf sayısı belli değil
- Uygulamaları:
  - verinin dağılımını anlama
  - başka veri madenciliği uygulamaları için ön hazırlık



<http://www3.itu.edu.tr/~sgunduz/courses/verimaden/>

4

## Demetleme Uygulamaları

- Örüntü tanıma
- Görüntü işleme
- Ekonomi
- Aykırılıkları belirleme
- WWW
  - Doküman demetleme
  - Kullanıcı davranışlarını demetleme
  - Kullanıcıları demetleme
- Diğer veri madenciliği algoritmaları için bir ön işleme adımı
  - Veri azaltma – demet içindeki nesnelerin temsil edilmesi için demet merkezlerinin kullanılması

<http://www3.itu.edu.tr/~sgunduz/courses/verimaden/>

5

## Veri Madenciliğinde Demetleme

- Ölçeklenebilirlik
- Farklı tipteki niteliklerden oluşan nesneleri demetleme
- Farklı şekillerdeki demetleri oluşturabilme
- En az sayıda giriş parametresi gereksinimi
- Hatalı veriler ve aykırılıklardan en az etkilenme
- Model oluşturma sırasında örneklerin sırasından etkilenmeme
- Çok boyutlu veriler üzerinde çalışma
- Kullanıcıların kısıtlarını göz önünde bulundurma
- Sonucun yorumlanabilir ve anlaşılabilir olması

<http://www3.itu.edu.tr/~sgunduz/courses/verimaden/>

6

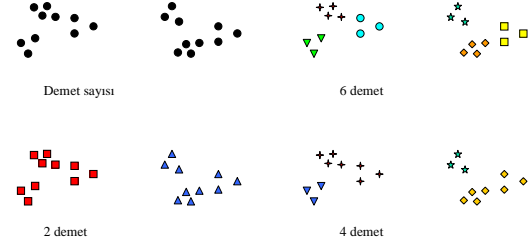
## İyi Demetleme

- İyi demetleme yöntemiyle elde edilen demetlerin özellikleri
  - aynı demet içindeki nesneler arası benzerlik fazla
  - farklı demetlerde bulunan nesneler arası benzerlik az
- Oluşan demetlerin kalitesi seçilen benzerlik ölçütüne ve bu ölçütün gerçekleştirilmesine bağlı
  - Uzaklık / Benzerlik nesnelerin nitelik tipine göre değişir
    - Nesneler arası benzerlik:  $s(i,j)$
    - Nesneler arası uzaklık:  $d(i,j) = 1 - s(i,j)$
- İyi bir demetleme yöntemi veri içinde gizlenmiş örüntüleri bulabilmeli
- Veriyi gruplama için uygun demetleme kriteri bulunmalı
  - demetleme = aynı demetteki nesneler arası benzerliği enbüyüten, farklı demetlerdeki nesneler arası benzerliği enküçüten fonksiyon
- Demetleme sonucunun kalitesi seçilen demetlerin şekline ve temsil edilme yöntemine bağlı

<http://www3.itu.edu.tr/~sgunduz/courses/verimaden/>

7

## Farklı Demetler



<http://www3.itu.edu.tr/~sgunduz/courses/verimaden/>

8

## Temel Demetleme Yaklaşımları

- Bölünmeli yöntemler: Veriyi bölerek, her grubu belirlenmiş bir kriterle göre değerlendirir
- Hiyerarşik yöntemler: Veri kümelerini (ya da nesneleri) önceden belirlenmiş bir kriterle göre hiyerarşik olarak ayırır
- Yoğunluk tabanlı yöntemler: Nesnelerin yoğunluğuna göre demetleri oluşturur
- Model tabanlı yöntemler: Her demetin bir modele uyduğu varsayılır. Amaç bu modellere uyan verileri gruplamak

<http://www3.itu.edu.tr/~sgunduz/courses/verimaden/>

9

## Konular

- Demetleme işlemleri
  - Demetleme tanımı
  - Demetleme uygulamaları
- Demetleme Yöntemleri
  - Bölünmeli Yöntemler
    - K-means demetleme yöntemi
    - K-medoids demetleme yöntemi
  - Hiyerarşik Yöntemler
  - Yoğunluk Tabanlı Yöntemler
  - Model Tabanlı Yöntemler

<http://www3.itu.edu.tr/~sgunduz/courses/verimaden/>

10

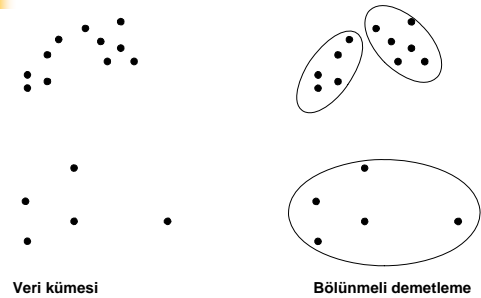
## Bölünmeli Yöntemler

- Amaç:  $n$  nesneden oluşan bir veri kümesini ( $D$ )  $k$  ( $k \leq n$ ) demete ayırmak
  - her demette en az bir nesne bulunmalı
  - her nesne sadece bir demette bulunmalı
- Yöntem: Demetleme kriterini enbüyütücek şekilde  $D$  veri kümesi  $k$  gruba ayırma
  - Global çözüm: Mümkün olan tüm gruplamaları yaparak en iyisini seçme (NP karmaşık)
  - Sezgisel çözüm: k-means ve k-medoids
    - k-means (MacQueen'67): Her demet kendi merkezi ile temsil edilir
    - k-medoids veya PAM (Partition around medoids) (Kaufman & Rousseeuw'87): Her demet, demette bulunan bir nesne ile temsil edilir

<http://www3.itu.edu.tr/~sgunduz/courses/verimaden/>

11

## Bölünmeli Demetleme



<http://www3.itu.edu.tr/~sgunduz/courses/verimaden/>

12

## K-means Demetleme

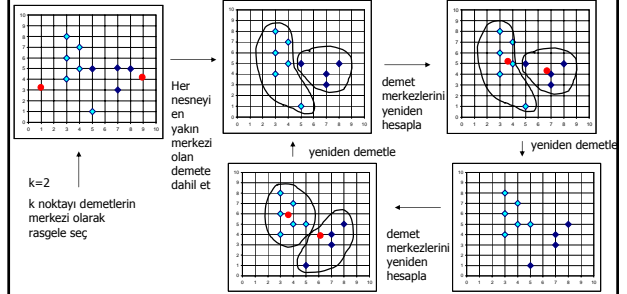
- Bilinen bir  $k$  değeri için k-means demetleme algoritmasının 4 aşaması vardır:
  - Veri kümesi  $k$  altkümeye ayrılır (her demet bir altküme)
  - Her demetin ortalaması hesaplanır: merkez nokta (demetteki nesnelerin niteliklerinin ortalaması)
  - Her nesne en yakın merkez noktanın olduğu demete dahil edilir
  - Nesnelerin demetlenmesinde değişiklik olmayana kadar adım 2'ye geri dönlür.

<http://www3.itu.edu.tr/~sgunduz/courses/verimaden/>

13

## K-means Demetleme Yöntemi

### Örnek



<http://www3.itu.edu.tr/~sgunduz/courses/verimaden/>

14

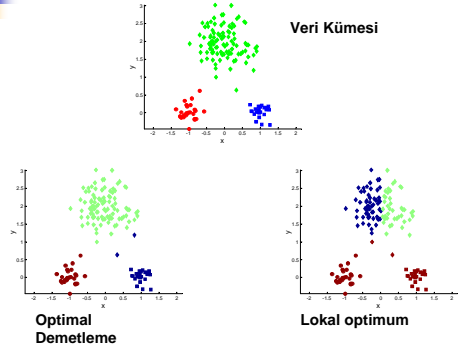
## k-means Demetleme Yöntemi

- Demet sayısının belirlenmesi gerekir
  - Başlangıçta demet merkezleri rasgele belirlenir
    - Her uygulamada farklı demetler oluşabilir
  - Benzerlik Öklid uzaklığı, kosinüs benzerliği gibi yöntemlerle ölçülebilir
  - Az sayıda tekrarda demetler oluşur
    - Yakınsama koşulu çoğunlukla az sayıda nesnenin demet değiştirmesi şekline dönüştürülür
  - Karmaşıklık:
    - Yer karmaşıklığı -  $O((n+k)d)$
    - Zaman karmaşıklığı -  $O(knd)$
- $k$ : demet sayısı,  $t$ : tekrar sayısı,  $n$ : nesne sayısı,  $d$ : nitelik sayısı

<http://www3.itu.edu.tr/~sgunduz/courses/verimaden/>

15

## K-Means: İki Farklı Demetleme



<http://www3.itu.edu.tr/~sgunduz/courses/verimaden/>

16

## K-Means Demetleme Yöntemini Değerlendirme

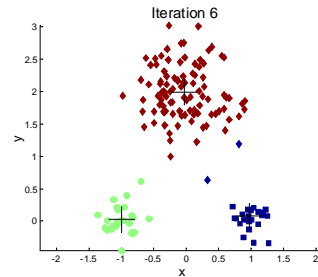
- Yaygın olarak kullanılan yöntem hataların karelerinin toplamı (Sum of Squared Error SSE)
  - Nesnelerin bulundukları demetin merkez noktalarına olan uzaklıklarının karelerinin toplamı
$$SSE = \sum_{i=1}^k \sum_{x \in C_i} dist^2(m_i, x)$$

$x$ :  $C_i$  demetinde bulunan bir nesne,  $m_i$ :  $C_i$  demetinin merkez noktası
- Hataların karelerinin toplamını azaltmak için  $k$  demet sayısı artırılabilir
  - Küçük  $k$  ile iyi bir demetleme, büyük  $k$  ile kötü bir demetlemeden daha az SSE değerine sahip olabilir.
- Başlangıç için farklı merkez noktaları seçerek farklı demetlemeler oluşturulur
- En az SSE değerini sahip olan demetleme seçilir

<http://www3.itu.edu.tr/~sgunduz/courses/verimaden/>

17

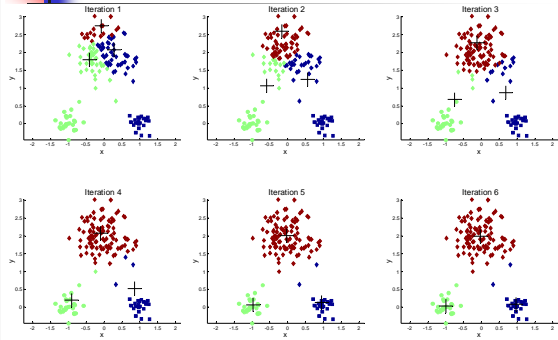
## Merkez Noktaların Seçimi



<http://www3.itu.edu.tr/~sgunduz/courses/verimaden/>

18

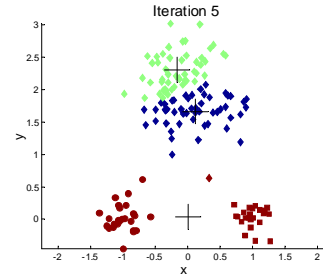
## Merkez Noktaların Seçimi



<http://www3.itu.edu.tr/~sgunduz/courses/verimaden/>

19

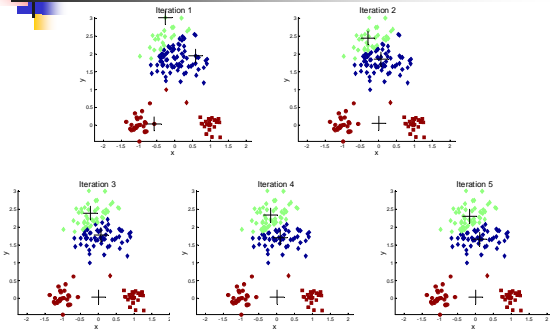
## Merkez Noktaların Seçimi



<http://www3.itu.edu.tr/~sgunduz/courses/verimaden/>

20

## Merkez Noktaların Seçimi



<http://www3.itu.edu.tr/~sgunduz/courses/verimaden/>

21

## K-Means Demetleme Çeşitleri

- K-Means demetlemeye başlamadan önce yapılanlar
  - Veri kümesini örnekleyerek hiyerarşik demetleme yap. Oluşan  $k$  demetin ortalamasını başlangıç için merkez nokta seç
  - Başlangıçta  $K$ 'dan fazla merkez nokta seç. Daha sonra bunlar arasından  $k$  tane seç.
- K-Means demetleme işlemi sonrasında yapılanlar
  - Küçük demetleri en yakın başka demetlere dahil et
  - En büyük toplam karesel hataya sahip olan demeti böl
  - Merkez noktaları birbirine en yakın demetleri birleştir
  - Toplam karesel hatada en az artışa neden olacak iki demeti birleştir

<http://www3.itu.edu.tr/~sgunduz/courses/verimaden/>

22

## K-Means Demetleme Algoritmasının Özellikleri

- Gerçeklemesi kolay
- Karmaşıklığı diğer demetleme yöntemlerine göre az
- K-Means algoritması bazı durumlarda iyi sonuç vermeyebilir
  - Veri grupları farklı boyutlarda ise
  - Veri gruplarının yoğunlukları farklı ise
  - Veri gruplarının şekli küresel değilse
  - Veri içinde ayırıcılıklar varsa

<http://www3.itu.edu.tr/~sgunduz/courses/verimaden/>

23

## Konular

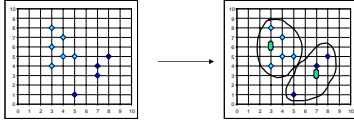
- Demetleme işlemleri
  - Demetleme tanımı
  - Demetleme uygulamaları
- Demetleme Yöntemleri
  - Bölünmeli Yöntemler
    - K-means demetleme yöntemi
    - K-medoids demetleme yöntemi
  - Hiyerarşik Yöntemler
  - Yoğunluk Tabanlı Yöntemler
  - Model Tabanlı Yöntemler

<http://www3.itu.edu.tr/~sgunduz/courses/verimaden/>

24

## K-Medoids Demetleme Yöntemi

- Her demeti temsil etmek için demet içinde orta nokta olan nesne seçilir.
  - 1, 3, 5, 7, 9 ortalama: **5**
  - 1, 3, 5, 7, 1009 ortalama: **205**
  - 1, 3, 5, 7, 1009 orta nokta: **5**



<http://www3.itu.edu.tr/~sgunduz/courses/verimaden/>

25

## K-Medoids Demetleme Yöntemi

- PAM (Partitioning Around Medoids 1987)
    - Başlangıçta  $k$  adet nesne demetleri temsil etmek üzere rasgele seçilir  $x_{jk}$
    - Kalan nesneler en yakın merkez nesnenin bulunduğu demete dahil edilir
    - Merkez nesne olmayan rasgele bir nesne seçilir  $x_{rk}$
    - $x_{rk}$  merkez nesne olursa toplam karesel hatanın ne kadar değiştiğini bulunur
- $$TC_{ik} = \sum_{j=1}^{n_i} (x_{ik} - x_{jk})^2 - \sum_{j=1}^{n_i} (x_{ik} - x_{jk})^2$$
- $n_i$ :  $k$  demeti içindeki nesne sayısı  
 $x_{jk}$ :  $k$  demeti içindeki  $j$ . nesne
- $TC_{ik} < 0$  ise  $O_{rk}$  merkez nesne olarak atanır.
  - Demetlerde değişiklik oluşmayana kadar 3. adıma geri gidilir.
- Küçük veri kümeleri için iyi sonuç verebilir, ancak büyük veri kümeleri için uygun değil
  - CLARA (Kaufmann & Rousseeuw, 1990)
  - CLARANS (Ng & Han, 1994)

<http://www3.itu.edu.tr/~sgunduz/courses/verimaden/>

26

## Konular

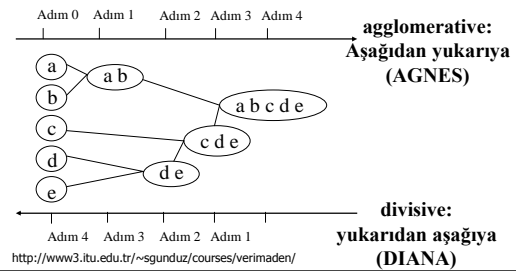
- Demetleme işlemleri
  - Demetleme tanımı
  - Demetleme uygulamaları
- Demetleme Yöntemleri
  - Bölünmeli Yöntemler
  - Hiyerarşik Yöntemler
  - Yoğunluk Tabanlı Yöntemler
  - Model Tabanlı Yöntemler

<http://www3.itu.edu.tr/~sgunduz/courses/verimaden/>

27

## Hiyerarşik Demetleme

- Demet sayısının belirlenmesine gerek yok
  - Sonlanma kriteri belirlenmesi gerekiyor

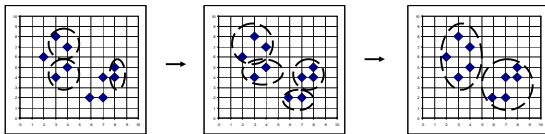


<http://www3.itu.edu.tr/~sgunduz/courses/verimaden/>

28

## Hiyerarşik Yöntemler

- AGNES (AGglomerative NESTing):
  - Kaufmann ve Rousseeuw tarafından 1990 yılında önerilmiştir.
  - Birinci adımda her nesne bir demet oluşturur.
  - Aralarında en az uzaklık bulunan demetler her adımda birleştirilir.
  - Bütün nesneler tek bir demet içinde kalana kadar ya da istenen sayıda demet elde edene kadar birleştirme işlemi devam eder.

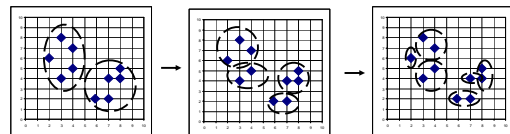


<http://www3.itu.edu.tr/~sgunduz/courses/verimaden/>

29

## Hiyerarşik Yöntemler

- DIANA (Dİvisive ANALysis):
  - Kaufmann ve Rousseeuw tarafından 1990 yılında önerilmiştir.
  - AGNES'in yaptığı işlemlerin tersini yapar.
  - En sonunda her nesne bir demet oluşturur.
  - Her nesne ayrı bir demet oluşturma ya da istenilen demet sayısı elde edene kadar ayrılma işlemi devam eder.

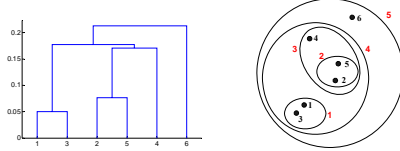


<http://www3.itu.edu.tr/~sgunduz/courses/verimaden/>

30

## Hiyerarşik Demetleme

- Dendrogram: Demetler hiyerarşik olarak ağaç yapısı şeklinde görüntülenebilir
- Ara düğümler çocuk düğümlerdeki demetlerin birleşmesiyle elde edilir
  - Kök: bütün nesnelerden oluşan tek demet
  - Yapraklar: bir nesneden oluşan demetler
- Dendrogram istenen seviyede kesilerek demetler elde edilir



<http://www3.itu.edu.tr/~sgunduz/courses/verimaden/>

31

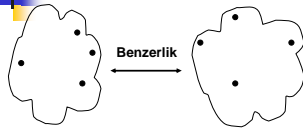
## Aşağıdan Yukarıya Demetleme

- Algoritma
  1. Uzaklık matrisini hesapla
  2. Her nesne bir demet
  3. Tekrarla
  4. En yakın iki demeti birleştir
  5. Uzaklık matrisini yeniden hesapla
  6. Sonlanma: Tek bir demet kalana kadar
- Uzaklık matrisini hesaplarırken farklı yöntemler farklı demetleme sonuçlarına neden olurlar

<http://www3.itu.edu.tr/~sgunduz/courses/verimaden/>

32

## Demetler Arası Uzaklık



- MIN (Tek Bağ)
- MAX (Tam Bağ)
- Ortalama
- Merkezler arası uzaklık

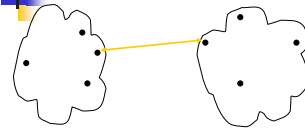
	p1	p2	p3	p4	p5	...
p1						
p2						
p3						
p4						
p5						
.						
.						

Uzaklık Matrisi

<http://www3.itu.edu.tr/~sgunduz/courses/verimaden/>

33

## Demetler Arası Uzaklık



- MIN (Tek Bağ)
- MAX (Tam Bağ)
- Ortalama
- Merkezler arası uzaklık

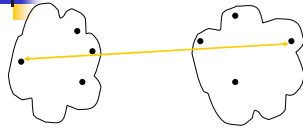
	p1	p2	p3	p4	p5	...
p1						
p2						
p3						
p4						
p5						
.						
.						

Uzaklık Matrisi

<http://www3.itu.edu.tr/~sgunduz/courses/verimaden/>

34

## Demetler Arası Uzaklık



- MIN (Tek Bağ)
- MAX (Tam Bağ)
- Ortalama
- Merkezler arası uzaklık

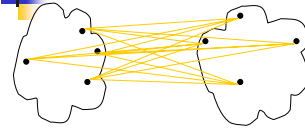
	p1	p2	p3	p4	p5	...
p1						
p2						
p3						
p4						
p5						
.						
.						

Uzaklık Matrisi

<http://www3.itu.edu.tr/~sgunduz/courses/verimaden/>

35

## Demetler Arası Uzaklık



- MIN (Tek Bağ)
- MAX (Tam Bağ)
- Ortalama
- Merkezler arası uzaklık

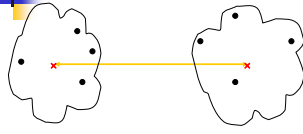
	p1	p2	p3	p4	p5	...
p1						
p2						
p3						
p4						
p5						
.						
.						

Uzaklık Matrisi

<http://www3.itu.edu.tr/~sgunduz/courses/verimaden/>

36

## Demetler Arası Uzaklık



- MIN (Tek Bağ)
- MAX (Tam Bağ)
- Ortalama
- Merkezler arası uzaklık

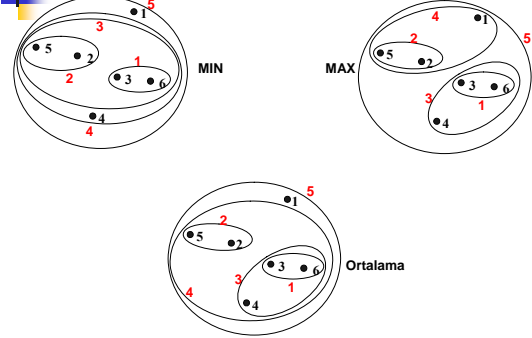
	p1	p2	p3	p4	p5	...
p1						
p2						
p3						
p4						
p5						
.						
.						

Uzaklık Matrisi

<http://www3.itu.edu.tr/~sgunduz/courses/verimaden/>

37

## Farklı Uzaklık Yöntemlerinin Etkisi



<http://www3.itu.edu.tr/~sgunduz/courses/verimaden/>

38

## Hiyerarşik Demetleme Yöntemlerinin Özellikleri

- Demetleme kriteri yok
- Demet sayılarının belirlenmesine gerek yok
- Aykırılıklardan ve hatalı verilerden etkilenir
- Farklı boyuttaki demetleri oluşturmak problemli olabilir
- Yer karmaşıklığı –  $O(n^2)$
- Zaman karmaşıklığı –  $O(n^2 \log n)$
- n : nesne sayısı

<http://www3.itu.edu.tr/~sgunduz/courses/verimaden/>

39

## Konular

- Demetleme işlemleri
  - Demetleme tanımı
  - Demetleme uygulamaları
- Demetleme Yöntemleri
  - Bölünmeli Yöntemler
  - Hiyerarşik Yöntemler
  - Yoğunluk Tabanlı Yöntemler
  - Model Tabanlı Yöntemler

<http://www3.itu.edu.tr/~sgunduz/courses/verimaden/>

40

## Yoğunluk Tabanlı Yöntemler

- Demetleme nesnelerin yoğunluğuna göre yapılır.
- Başlıca özellikleri:
  - Rasgele şekillerde demetler üretilebilir.
  - Aykırı nesnelerden etkilenmez.
  - Algoritmanın son bulması için yoğunluk parametresinin verilmesi gerekir.
- Başlıca yoğunluk tabanlı yöntemler:
  - DBSCAN: Ester, et al. (KDD'96)
  - OPTICS: Ankerst, et al (SIGMOD'99).
  - DENCLUE: Hinneburg & D. Keim (KDD'98)
  - CLIQUE: Agrawal, et al. (SIGMOD'98)

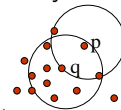
<http://www3.itu.edu.tr/~sgunduz/courses/verimaden/>

41

## DBSCAN

- İki parametre:
  - $Eps$ : En büyük komşuluk yarıçapı
  - $MinPts$ :  $Eps$  yarıçaplı komşuluk bölgesinde bulunan en az nesne sayısı
- $N_{eps}(p)$ :  $\{q \in D \mid d(p,q) \leq Eps\}$
- Doğrudan erişilebilir nesne:  $Eps$  ve  $MinPts$  koşulları altında bir  $q$  nesnesinin doğrudan erişilebilir bir  $p$  nesnesi şu şartları sağlar:
  - $p \in N_{eps}(q)$
  - $q$  nesnesinin çekirdek nesne koşulunu sağlaması

$$N_{eps}(q) \geq MinPts$$



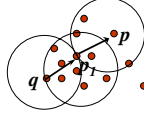
MinPts = 5  
Eps = 1 cm

<http://www3.itu.edu.tr/~sgunduz/courses/verimaden/>

42

## DBSCAN

- Erişilebilir nesne:
  - $Eps$  ve  $MinPts$  koşulları altında  $q$  nesnesinin erişilebilir bir  $p$  nesnesi olması için:
    - $p_1, p_2, \dots, p_n$  nesne zinciri olması,
    - $p_1 = q, p_n = p$ ,
    - $p_i$  nesnesinin doğrudan erişilebilir nesnesi:  $p_{i+1}$
- Yoğunluk bağlantılı Nesne:
  - $Eps$  ve  $MinPts$  koşulları altında  $q$  nesnesinin yoğunluk bağlantılı nesnesi  $p$  şu koşulları sağlar:
    - $p$  ve  $q$  nesneleri  $Eps$  ve  $MinPts$  koşulları altında bir  $o$  nesnesinin erişilebilir nesnesidir.



<http://www3.itu.edu.tr/~sgunduz/courses/verimaden/>

43

## Yoğunluk Tabanlı Yöntemler: DBSCAN

- Veri tabanındaki her nesnenin  $Eps$  yarıçaplı komşuluk bölgesi araştırılır.
- Bu bölgede  $MinPts$ 'den daha fazla nesne bulunan  $p$  nesnesi çekirdek nesne olacak şekilde demetler oluşturulur.
- Çekirdek nesnelerin doğrudan erişilebilir nesneleri bulunur.
- Yoğunluk bağlantılı demetler birleştirilir.
- Hiçbir yeni nesne bir demete eklenmezse işlem sona erer.
- Yer karmaşıklığı –  $O(n)$
- Zaman karmaşıklığı –  $O(n \log n)$   
n: nesne sayısı

<http://www3.itu.edu.tr/~sgunduz/courses/verimaden/>

44

## Konular

- Demetleme işlemleri
  - Demetleme tanımı
  - Demetleme uygulamaları
- Demetleme Yöntemleri
  - Bölünmeli Yöntemler
  - Hiyerarşik Yöntemler
  - Yoğunluk Tabanlı Yöntemler
  - Model Tabanlı Yöntemler

<http://www3.itu.edu.tr/~sgunduz/courses/verimaden/>

45

## Model Tabanlı Demetleme Yöntemleri

- Veri kümesi için öngörülen matematiksel model en uygun hale getiriliyor.
- Verinin genel olarak belli olasılık dağılımlarının karışımından geldiği kabul edilir.
- Model tabanlı demetleme yöntemi
  - Modelin yapısının belirlenmesi
  - Modelin parametrelerinin belirlenmesi
- Örnek EM (Expectation Maximization) Algoritması

<http://www3.itu.edu.tr/~sgunduz/courses/verimaden/>

46

## Model Tabanlı Demetleme Yöntemleri

- İstatistiksel yaklaşım:
  - K nesneden oluşan bir veri kümesi  $D = \{x_1, x_2, \dots, x_K\}$
  - her  $x_i (i \in \{1, \dots, K\})$  nesnesi  $\theta$  parametre kümesiyle tanımlanan bir olasılık dağılımından oluşturulur.
  - Olasılık dağılımının,  $c_g \in \{c_1, c_2, \dots, c_G\}$  şeklinde  $G$  adet bileşeni vardır.
  - Her  $\theta_g, g \in \{1, \dots, G\}$  parametre kümesi  $g$  bileşeninin olasılık dağılımını belirleyen,  $\theta$  kümesinin ayrışık bir alt kümesidir.
  - Herhangi bir  $x_i$  nesnesi öncelikle,  $p(c_g | \theta) = \tau_g$  ( $\sum_g \tau_g = 1$  olacak şekilde) bileşen katsayısına (ya da bileşenin seçilme olasılığına) göre bir bileşene atanır.
  - Bu bileşen  $p(x_i | c_g; \theta_g)$  olasılık dağılımına göre  $x_i$  değişkenini oluşturur.
  - Böylece bir  $x_i$  nesnesinin bu model için olasılığı bütün bileşenlerin olasılıklarının toplamıyla ifade edilebilir:

$$p(x_i | \theta) = \sum_{g=1}^G p(c_g | \theta) p(x_i | c_g; \theta_g)$$

$$p(x_i | \theta) = \sum_{g=1}^G \tau_g p(x_i | c_g; \theta_g)$$

<http://www3.itu.edu.tr/~sgunduz/courses/verimaden/>

47

## Model Tabanlı Demetleme Problemi

- Model parametrelerinin belirlenmesi

- Maximum Likelihood (ML) yaklaşımı

$$\ell_{ML}(\Theta_1, \dots, \Theta_G; \tau_1, \dots, \tau_G | D) = \prod_{i=1}^K \sum_{g=1}^G \tau_g p(x_i | c_g; \theta_g)$$

- Maximum Aposteriori (MAP) yaklaşımı

$$\ell_{MAP}(\Theta_1, \dots, \Theta_G; \tau_1, \dots, \tau_G | D) = \prod_{i=1}^K \sum_{g=1}^G \frac{\tau_g p(x_i | c_g; \theta_g) p(\theta_g)}{p(D)}$$

- Uygulamada her ikisinin logaritması

$$L(\Theta_1, \dots, \Theta_G; \tau_1, \dots, \tau_G | D) = \sum_{i=1}^K \sum_{g=1}^G (\tau_g p(x_i | c_g; \theta_g))$$

$$L(\Theta_1, \dots, \Theta_G; \tau_1, \dots, \tau_G | D) = \sum_{i=1}^K \sum_{g=1}^G (\tau_g p(x_i | c_g; \theta_g)) + \ln p(\theta)$$

<http://www3.itu.edu.tr/~sgunduz/courses/verimaden/>

48



## EM Algoritması

- Veri kümesi:  $D=\{x_1, x_2, \dots, x_K\}$
- Gizli değişkenler  $H=\{z_1, z_2, \dots, z_K\}$  (her nesnenin hangi demete dahil olduğu bilgisi)
- Verinin eksik olduğu durumda, tam verinin beklenen değeri hesaplanır:

$$\begin{aligned} Q(\Theta, \Theta') &= E[L_{\Theta}(D, H | \Theta) | D, \Theta'] \\ &= \sum_{i=1}^K \sum_{g=1}^G p(c_g | x_i) [\ln p(x_i | c_g) + \ln \tau_g] \end{aligned}$$

- EM Algoritmasının adımları:
  - $\Theta'$  için başlangıç değerleri atama
  - (E) Expectation:  $Q(\Theta | \Theta')$  hesaplanması
  - (M) Maximization:  $\operatorname{argmax} Q(\Theta | \Theta')$

<http://www3.itu.edu.tr/~sgunduz/courses/verimaden/>

49