Veri Madenciliği

Ders Notları - 2

Verilerin Önişlenmesi

- Verilerin önişlenmesi nedenleri
- Veri temizleme
- Veri bütünleştirme ve dönüştürme
- Veri küçültme
- Ayrıklaştırma ve kavram hiyerarşisi

Verilerin çok boyutlu niteliği

- Verilerin projenin amacına uygunluk derecesini belirlemek için onların çeşitli boyutlarda değerlendirilmesi gerekmektedir:
  - kesinlik
  - tamlık
  - tutarlılık
  - zamanlama
  - güvenilirlik
  - Yorumlanabilirlik
  - Erişebilirlik
- Çoğu zaman çeşitli nedenlerden dolayı veriler bu boyutlardan bir veya birkaçı üzere gereken koşulları sağlamıyor. Bu durumda verilerin önişlenmesine ihtiyaç duyuluyor.

Verilerin Önişlenmesi nedenleri

- Kullanılmadan önce verilerin önişlenmesinin nedenleri:
  - Veriler tam değil: özelliklerin bazı değerlerinin bulunmaması
    - örneğin., maaşı=“ ”
  - Veriler gürültülüdür(parazit): hatalar veya sapmalar içerir
    - Örn., maaş=“-10”
  - Veriler tutarlı değil: Değişkenlerin değerleri arasında tutarsızlık bulunmaktadır
    - Yaş=“42” Doğum günü=“03/07/1997”
    - Bir sıralamada “1,2,3”, diğerinde “A, B, C”

Veriler neden «kirlidir»

- Verilerin tam olmamasının nedenleri:
  - Verilere erişilememesi
  - Verilere, toplandığı ve çözümlendiği zaman dilimlerinde farklı yanaşmalar(bazı verilerin değerini önemsememe)
  - insan/donanım/yazılım sorunları
- Gürültülü (düzgün olmayan) verilerin nedenleri
  - Veri toplama araçlarında hatalar
  - Veri girişinde insan veya bilgisayar hatası
  - Veri aktarımında hatalar
- Tutarsız verilerin nedenleri
  - Farklı veri kaynakları
  - İşlevsel bağlılıklarda yanlışlar (bağımlı değişkenlerin değerlerinin doğru hesaplanmaması)

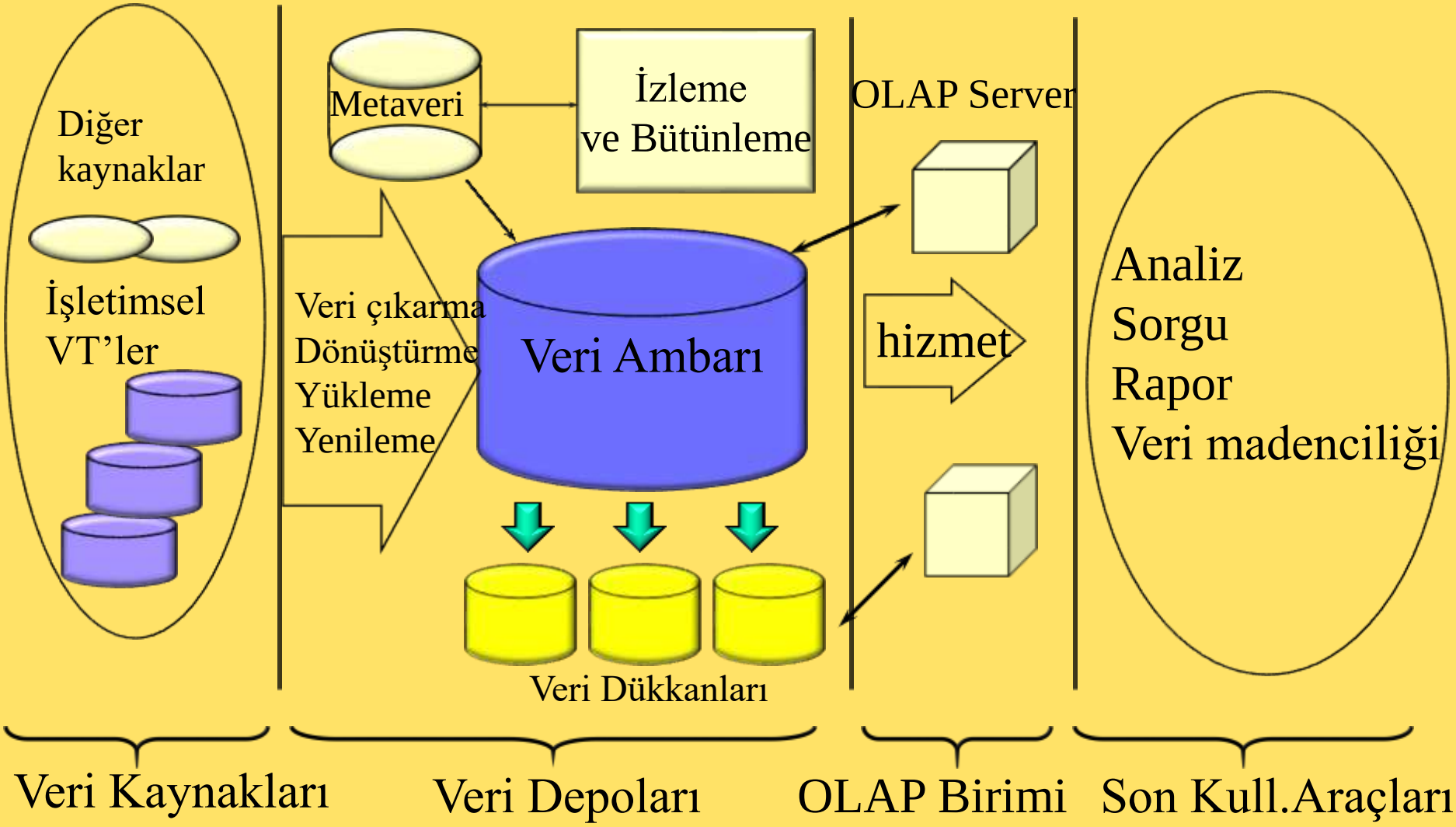
Veri kirliliği örneği-1

kapsam	sorun	Kirli veriler	sebep
özellik	Yanlış değer	Doğum_günü =30.13.1990	Değerler alan dışındadır
Kayıt	Özellikler arasında bağımlılığın yanlış olması	Yaş=42 Doğum_günü=12.02.1990	«yaş»la doğum günü değerleri tutarsızdır
Kayıt türü	Eşsizliğin bozulması	Pers1=(ad=«Ali Yavuz», pno=«123456» Pers1=(ad=«Metin SAĞLAM», pno=«123456»	Personel numarasının eşsiz olması koşulu bozulmuştur
kaynak	Erişimsel bütünlüğün bozulması	Pers1=(ad=«Metin SAĞLAM», şube_no=«123456»	«123456»no'lu şube tanımlanmamıştır

Veri kirliliği örneği-2

kapsam	sorun	Kirli veriler	sebep
özellik	Değer yoktur	Tel:=285218 163	Rakam eksiktir
özellik	Kelimenin yanlış yazılışı	Kent=«Trabzun»	Fonetik hata
özellik	yanlış alan değeri	Kent=«İtalya»	«Italya» «kent» alanına dahil değil
kayıt	Özellikler arası bağımlılığın bozulması	Kent=«Çanakkale»; plaka_no=19	«Çanakkale'nin plaka numarası 19 değil
Kayıt türü	Kelimelerin farklı dizilişi	Ad1 =«Kerim UĞUR» Ad2=«YILMAZ Temel»	Ad ve soyadların sıraları farklıdır
Kayıt türü	Kayıtlarda zıtlık	Pers1=(ad=«Ali Yavuz», doğum_tar=12.12.1995 Pers2=(ad=«Ali Yavuz», doğum_tar=10.09.1995	Aynı varlık farklı değerlerle tanımlanmıştır

Farklı veri kaynakları: Veri Ambarı mimarisi



Veri Ambarı Nedir?

- Veri tabanları ve diğer veri kaynaklarından yönetici sorunlarının çözümünde kullanılacak veriyi elde etmek için gerekli olan algoritmaları, araçları içeren sistemdir
- Yönetici verilerini sorgulama ve raporlama için kullanılmaktadır.
- Bir veri ambarı ilgili veriyi kolay, hızlı, ve doğru bicimde analiz etmek için gerekli işlemleri yerine getirir. Veri ambarı, işletimsel sistemlerdeki veriyi karar verme işlemi için uygun biçimde saklar.

Veri Önişlemenin önemi

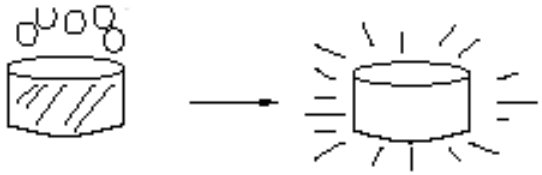
- Nitelikli veri olmadan nitelikli sonuç almak mümkün değil
  - Nitelikli karar, nitelikli verilere dayanmalıdır
  - Yönetici kararları için veri kaynaklarını sağlayan veri ambarları, nitelikli verilerin tutarlı bütünleşmesini gerektiriyor
- Veri çıkarma, temizleme ve dönüştürme veri ambarı oluşturma sürecinin esasıdır

Veri Önişlemenin temel meseleleri

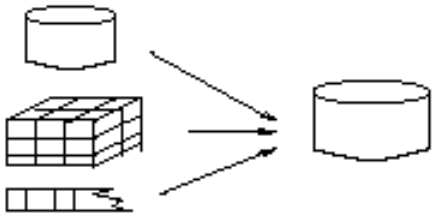
- Veri temizleme
  - Olmayan verilerin yerinin doldurulması, gürültülü verilerin düzlendirilmesi, sapmaların tanımlanması ve ya aradan kaldırılması, tutarsızlıkların çözülmesi
- Veri bütünleşmesi
  - Çoklu veri tabanlarının, dosyaların bütünleştirilmesi
- Veri dönüştürme
  - Normalleştirme ve bir yere yığma (aggregation)
- Veri küçültme
  - Aynı veya benzer sonuçlar almak koşuluyla verilerin ifade boyutlarının küçültülmesi
- Veri ayrıklaştırma
  - Özellikle, sayısal değerler için, önemli verileri dikkate almakla veri küçültme

Veri Önişleme biçimleri

veri temizleme



veri bütünleştirme



veri dönüştürme

-2, 32, 100, 59, 48 → -0.02, 0.32, 1.00, 0.59, 0.48

veri küçültme

Veri temizleme

Veri Temizleme

- Önemi:
  - "Veri temizleme, veri ambarları oluşturulmasında en esas sorunlardandır"
- Veri temizleme meseleleri
  - Eksik değerlerin yerinin doldurulması
  - Sapmaların tanımlanması ve gürültülü verilerin düzlendirilmesi
  - Tutarsız verilerin düzeltilmesi
  - Veri bütünleşmesi ile bağlı fazlalığın aradan götürülmesi

Veri temizleme- Eksik veriler

- Veri erişilemezdir:
  - Bazı özelliklerin değerleri kaydedilmemiştir; (satış verilerinde müşteri gelirleri gibi…)
- Veri eksikliğinin nedenleri:
  - Donanım hatası
  - Diğer kaydedilmiş verilerle tutarsızlık ve bu nedenle silinmesi
  - Doğru anlaşılmadığı için veri girilmemiştir
  - Veri girişi sürecinde bazı veriler önemsiz sayılarak girilmemiştir
  - Verinin oluşma veya değişme tarihi yoktur
- Eksik veriler karar alma zamanı gerekli olabilir.

Veri temizleme- Eksik verilerle işleme

Neler yapılabilir:

- Eksik veri olan satırı dikkate almamalı
- Veri değerini elle girmeli:

Değerleri

- Genel sabit gibi, örn. *"belli değil"* olarak *;*
- Özellik değerlerinin ortalaması olarak (*sınıfın gno'su bir öğrencinin gno'su olarak);*
- Aynı sınıfa ait tüm örneklerin özellik ortalaması olarak girmeli; (*öğrencinin matematik puanı belli değilse, gno'ları aynı olan öğrencilerin matematik puanlarının ortalaması olarak);*
- En ihtimal olunan değerin-Bayes formülü veya karar ağacı gibi çıkarıma yönelik değerin girilmesi; (*öğrencinin diğer notlarına bakmakla matematik notunun karar ağacı ile tahmin edilmesi)*

Veri temizleme- Gürültülü Değer

- Gürültü: ölçülen değişkende tesadüfü hata veya değişme
- Özellik değerlerinin düzgün olmaması nedenleri:
  - Veri toplama araçlarında hata
  - Veri girişi sorunları
  - Veri iletişimi sorunları
  - Teknoloji sınırlamalar
  - Dönüştürme zamanı tutarsızlık
- Veri temizlemesinde ortaya çıkan diğer sorunlar:
  - Tekrarlanan kayıtlar
  - Tam olmayan veriler
  - Tutarsız veriler

Veri temizleme- Gürültülü verilerle işleme

- Sepetlere ayırma-Binning
  - Verileri sıralamalı ve eşit sıklıklı sepetlere-bölümlere ayırmalı
  - Bölümler bölüm ortalamasına, bölüm medyanına, bölüm sınırlarına… göre düzlendirilir
- Regresyon
  - Regresyon fonksiyonları üzere düzlendirme
- Kümeleme-Clustering
  - Sapmaları bulma ve silme
- Bilgisayar ve insan gözlemlerinin birleştirilmesi
  - Kuşkulu değerleri bulma ve yoklama

Veri temizleme- Veri düzleştirme için sepetlere bölme yöntemleri-Binning Methods for Data

1. Verileri değerlerine göre sıralamalı: 4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34

2. Sıralanmış verileri eşit derinlikli (aynı sayıda elementlerden oluşan) sepetlere ayırmalı

- Sepet1: 4, 8, 9, 15
- Sepet2: 21, 21, 24, 25
- Sepet3: 26, 28, 29, 34

3. Verilerin değerini değiştirmeli (düzleştirme –(*smooth)* yapmalı)

* Bölüm ortalamasına göre düzleştirme
  - Sepet1: 9, 9, 9, 9
  - Sepet2: 23, 23, 23, 23
  - Sepet3: 29, 29, 29, 29
* Bölüm sınırlarına göre düzleştirme
  - Sepet1: 4, 4, 15, 15
  - Sepet2: 21, 21, 25, 25
  - Sepet3: 26, 26, 34, 34

Korelasyon

- **Korelasyon**, olasılık kuramı ve istatistikte iki rassal değişken arasındaki doğrusal ilişkinin yönünü ve gücünü belirtir

- Korelasyon katsayısı, bağımsız değişkenler arasındaki ilişkinin yönü ve büyüklüğünü belirten katsayıdır. Bu katsayı, (-1) ile (+1) arasında bir değer alır. Pozitif değerler doğru yönlü doğrusal ilişkiyi; negatif değerler ise ters yönlü bir doğrusal ilişkiyi belirtir. Korelasyon katsayısı 0 ise söz konusu değişkenler arasında doğrusal bir ilişki yoktur

- Korelasyon veya doğrusal ilişki nedensellik değildir.

Korelasyon ve nedensellik

- A ve B arasında korelasyon incelenince üç tür mümkün ilişki olabileceği görülür:

  A nedendir B sonuçtur;

  B nedendir A sonuçtur;

  C neden A sonuçtur VE C neden B sonuçtur.

  A ve B arasında görülen ilişkinin sebep-sonuç ilişkisi olması her zaman doğru olmayabilir. Bu sahte korelasyondur.

Sahte korelasyon örnekleri

- Bir sahil şehrinde aylık dondurma satışları ile aylık denizde boğulma sayıları yıl içinde birlikte artıp eksilime gösterip yakın pozitif korelasyon gösterirler. Bu demek değildir ki fazla dondurma fazla boğulmalara sebep-sonuç olmakta veya boğulmaların azalması dondurma satışlarına aksi tesirde bulunmaktadır. Her ikisi de mevsim değiştiği için aynı yönde değişik etki görmektedir.

- 1950lerden beri hava kirliği göstergeleri ile polise bildirilen hırsızlık olayları sayısı pozitif korelasyon göstermektedir. Bu demek değildir ki hava kirliği artışı hırsızlık olaylarının artışına; yahut hava kirliğinin artışı hırsızlık sayısı artışına neden olmuştur. Her iki değişken de hızlı şehirleşme dolayısı ile artış göstermektedir.

Korelasyon ilişki analizi (Sayısal Veriler)

- Korelasyon katsayısı

$$r_{A,B} = \frac{\sum (A-\overline{A})(B-\overline{B})}{(n-1)\sigma_A \sigma_B} = \frac{\sum (AB) - n\overline{A}\overline{B}}{(n-1)\sigma_A \sigma_B}$$

n- satırlar sayısı, $\bar{A}$ ve $\bar{B}$ uygun olarak A ve B'nin ortalamaları, σ_A ve σ_B - A ve B'nin standart sapmaları, Σ(AB) - AB çapraz çarpımının toplamıdır.

- Eğer $r_{A,B} > 0$ ise, A ve B – pozitif ilişkilidir (A'nın değeri yükseldikçe B de yükseliyor). Ne kadar yüksek ise, ilişki o kadar güçlüdür
- $r_{A,B} = 0$: bağımsız; $r_{A,B} < 0$: negatif ilişkili

Korelasyon Analizi (Kategorik veriler)

- X^2 (chi-square) denemesi

$$\chi^2 = \sum \frac{(Gözlenen - Beklenen)^2}{Beklenen}$$

- X^2 değeri büyük olması , değişkenlerin yakınlığının az olmasını gösteriyor

Korelasyon nedensellik anlamına gelmez

- Kentteki hastaneler sayısı ve araba hırsızlığı sayısı ilişkilidir.
- Her ikisi nedensel olarak üçüncü bir değişkene- nüfuz sayısına bağlıdır

Regresyon Analizi

- Regresyon analizi, bilinen bulgulardan, bilinmeyen veya gelecekteki olaylarla ilgili tahminler yapılmasına izin verir. Regresyon, bağımlı ve bağımsız değişken(ler) arasındaki ilişkiyi ve doğrusal eğri kavramını kullanarak, bir tahmin eşitliği geliştirir.
- Bağımlı Değişken (y); Bağımlı değişken, regresyon modelinde açıklanan ya da tahmin edilen değişkendir.
- Bağımsız Değişken (x); Bağımsız değişken, regresyon modelinde açıklayıcı değişken olup; bağımlı değişkenin değerini tahmin etmek için kullanılır.
- Değişkenler arasında doğrusal ilişki olabileceği gibi, doğrusal olmayan bir ilişki de olabilir.

Regresyon Analizi (devamı)

- Bağımlı değişken ile bağımsız değişken arasındaki doğrusal ilişkiyi açıklayan tek değişkenli regresyon modeli aşağıdaki gibidir:

$$y=ax+b$$

- Burada

y = Bağımlı değişkenin değeri

a = Regresyon doğrusunun kesişim değeri (Sabit değer)

b = Regresyon doğrusunun eğimi

x = Bağımsız değişkenin değerini göstermektedir

Veri temizleme- Regresyon Analizi -örnek

- Kardiyoloji kliniğine başvuran erkek hastalar üzerinde yapılan bir araştırmada, yaş(x) ve kolesterol(y) değişkeni arasındaki korelasyondan yola çıkılarak kurulan regresyon modeli aşağıdaki gibi elde edilmiştir:
- Bu modele göre, yaştaki bir birimlik artışın, kolesterol değerinde 0.326 birimlik bir artışa neden olacağı, yeni doğan bir erkeğin (X=0) kolesterol değerinin ise 3.42 olacağı söylenebilir.
- Kurulan bu modele göre, 50 yaşında bir erkeğin kolesterol değerinin ne kadar olacağını tahmin edebiliriz
- X=50 için

50 yaşında bir erkeğin kolesterol değerinin 19.52 olacağı söylenebilir.

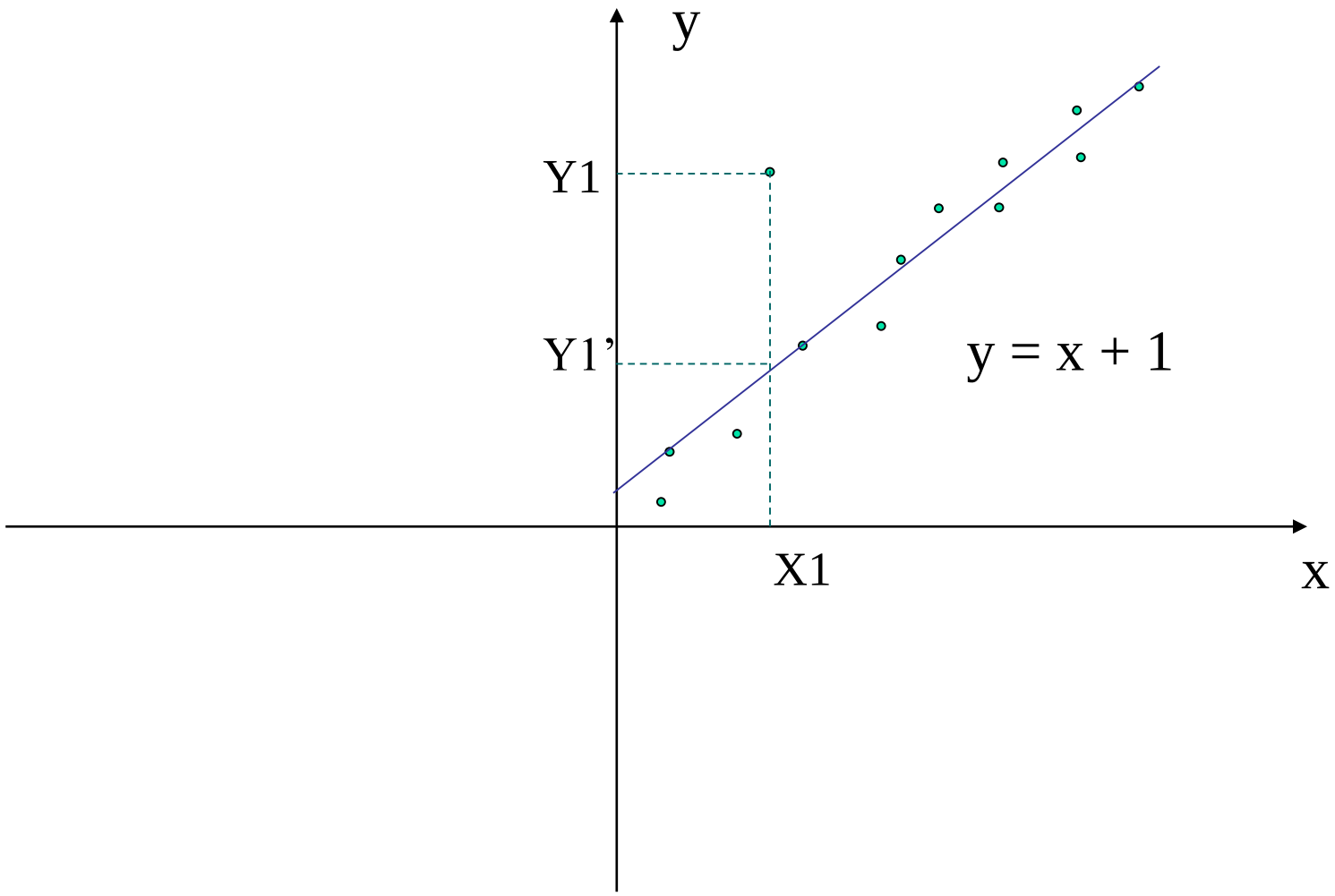
Korelasyon Analizi ve Regresyon Analizi Arasındaki fark

Korelasyon Analizi; iki veya daha çok değişken arasında ilişkinin varlığını, ilişki varsa yönünü ve gücünü inceler.

Regresyon Analizi; değişkenlerden birisi belirli bir birim değiştiği zaman, diğer değişkenlerin nasıl bir tepki verdiğini inceler.

İkisi arasındaki fark; korelasyon analizinde değişkenler arası ilişkiler incelenirken, diğer yanda regresyon analizinde ise; bir değişkenin değişiminde diğer değişkenlerin izlediği yol incelenir.

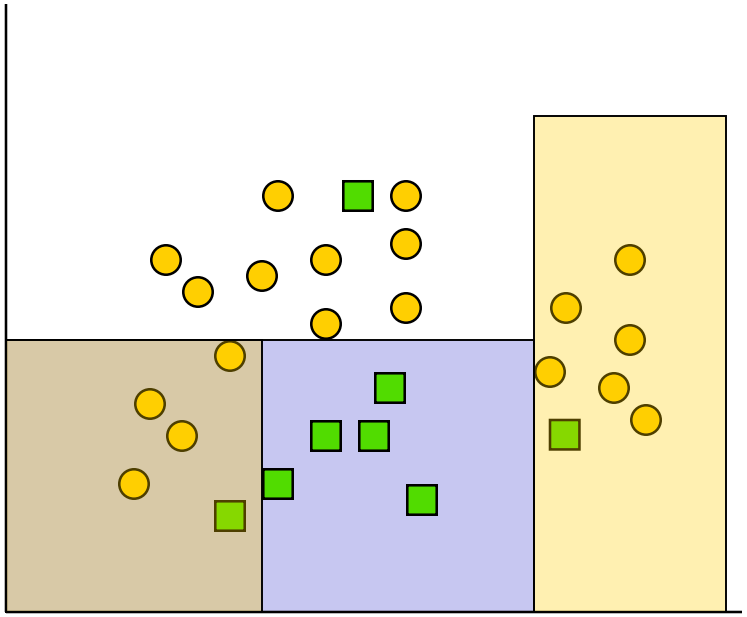
Veri temizleme- Regresyon-doğrusal ilişki



Sınıflandırma

Sınıflandırma veya Danışmanlı öğrenme:

Önceden etiketlenmiş (sınıflandırılmış) örnekler esasında yeni örneğin sınıfının belirlenmesi

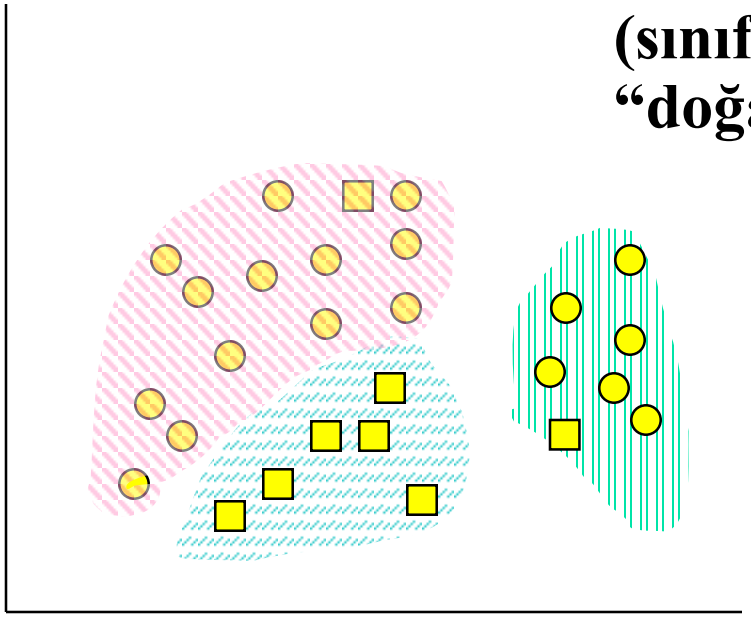


Sınıflar (dörtgenler) dışındaki veri, benzer (yakın) özellikleri bulunan sınıfa dahil edilir

Kümeleme

Kümeleme veya Danışmansız öğrenme:

Etiketlenmemiş (sınıflandırılmamış) verilerin "doğal" gruplaştırılması



Benzer(yakın) veriler küme oluşturuyor

Verilerin önişlenilmesi

- Veri Bütünleme ve Dönüştürme

Veri Bütünlemede fazlalığın aradan kaldırılması

- Çoklu veritabanlarının bütünleşmesi zamanı veri fazlalığı ortaya çıkıyor
  - *Nesne tanımlanması*: Aynı nesne veya özellik farklı veri tabanlarında farklı adlar taşımaktadır
  - *Alınma veriler:* Bir tablodaki özellik değeri, diğer bir tablodaki özellik değerlerinden alınabilir.
- Fazla (önemsiz) özelliklerin korelasyon analiz yöntemleriyle silinmesi mümkündür
- Farklı kaynaklardan alınmış verilerin bütünleştirilmesi sürecine özenli yaklaşımla veri fazlalığını ve tutarsızlığı azaltmak/küçültmek mümkündür. Bununla da gereken veriyi bulma hızı ve kalitesi yükselmiş olur.

Veri Bütünleme- Data Integration

- Veri bütünleme:
  - Pek çok kaynaktan verilerin bir depoda tutarlı biçimde birleşmesi
- Bütünleşme şeması: örn., A.müşt-id ≡ B.müşt-#
  - Farklı kaynaklardaki metaverilerin bütünleşmesi
- Varlık tanımlama sorunu:
  - Çoklu veri kaynaklarından gerçek dünya varlıklarının tanımlanması, örn., Bill Clinton = William Clinton
- Veri değerleri tutarsızlıklarını bulma ve çözme
  - Aynı gerçek dünya varlığı için , farklı kaynaklardan alınan özellik değerleri farklı olabilir
  - Mümkün nedenler: farklı sunumlar; farklı ölçekler, örn., metrik ve İngiliz ölçüm birimleri

Veri Bütünleme örneği

Customer (source 1)

CID	*Name*	*Street*	*City*	*Sex*
11	Kristen Smith	2 Hurley Pl	South Fork, MN 48503	0
24	Christian Smith	Hurley St 2	S Fork MN	1

Client (source 2)

Cno	*LastName*	*FirstName*	*Gender*	*Address*	*Phone/Fax*
24	Smith	Christoph	M	23 Harley St, Chicago IL, 60633-2394	333-222-6542 / 333-222-6599
493	Smith	Kris L.	F	2 Hurley Place, South Fork MN, 48503-5998	444-555-6666

Customers (integrated target with cleaned data)

No	*LName*	*FName*	*Gender*	*Street*	*City*	*State*	*ZIP*	*Phone*	*Fax*	*CID*	*Cno*
1	Smith	Kristen L.	F	2 Hurley Place	South Fork	MN	48503-5998	444-555-6666		11	493
2	Smith	Christian	M	2 Hurley Place	South Fork	MN	48503-5998			24	
3	Smith	Christoph	M	23 Harley Street	Chicago	IL	60633-2394	333-222-6542	333-222-6599		24

Figure 3. Examples of multi-source problems at schema and instance level

Müşteriler hakkında bilgiler iki farklı kaynaktan (customer ve client tablolarından) alınmıştır. Customers tablosu bu tablolardaki verileri temizlemekle alınmıştır.

Veri Dönüştürme

- düzleşdirme: verilerdeki gürültüleri silmek
- Bir yere toplama (Aggregation): verileri özetleme
- Genelleştirme: kavram hiyerarşisi
- Normalleştirme: değerin belirtilen aralık içine düşmesi için ölçekleme yapılması
  - min-max normalleştirme
  - z-score normalleştirme
  - Onluk ölçekte normalleştirme

Veri dönüştürme

- Min-max normalleştirme: [new_min$_A$, new_max$_A$]

$$v' = \frac{v - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A$$

  - Örnek: \$12,000- \$98,000 aralığındaki gelirleri [0.0, 1.0] aralığında normalleştirmek gerekmektedir. Varsayalım ki, gelir \$73,600 değerindedir. O zaman

$$\frac{73{,}600 - 12{,}000}{98{,}000 - 12{,}000}(1.0 - 0) + 0 = 0.716$$

Veri dönüştürme

- Z-score normalleştirme (μ: ortalama, σ: standard sapma):

$$v' = \frac{v - \mu_A}{\sigma_A}$$

  - Örnek: μ = 54,000, σ = 16,000. O zaman $\frac{73{,}600 - 54{,}000}{16{,}000} = 1.225$

- Onluk ölçekte normalleştirme

$$v' = \frac{v}{10^j} \quad ; j - \text{Max}(|v'|) < 1 \text{ yapan en küçük tam sayıdır}$$

Örnek: X özelliğinin değeri -500 - 45 aralığındadır. X'in en büyük mutlak değeri=500. Onluk ölçekte normalleştirmek için her değer 1000'e (j=3) bölünmelidir. Bizim örnekte -500 - 0.5'e dönüştürülecek. 45 ise 0.045 olacak

Dönüştürme: İkiliden sayısala

- İkili alan
  - Cinsiyet=M, F
- 0,1 değerli alana dönüştürme
  - Cinsiyet = M → Cinsiyet_0_1 = 0
  - Cinsiyet = F → Cinsiyet_0_1 = 1

Dönüştürme: Sıralıdan sayısala

- Sıralı özellikler, doğal sıralamayı sağlayan sayılara dönüştürüle biler:
  - A → 4.0
  - A- → 3.7
  - B+ → 3.3
  - B → 3.0
- Doğal sıralama, anlamsal karşılaştırma yapmak için önemlidir

Verilerin Önişlenmesi

- Veri küçültme

Veri Küçültme Stratejileri

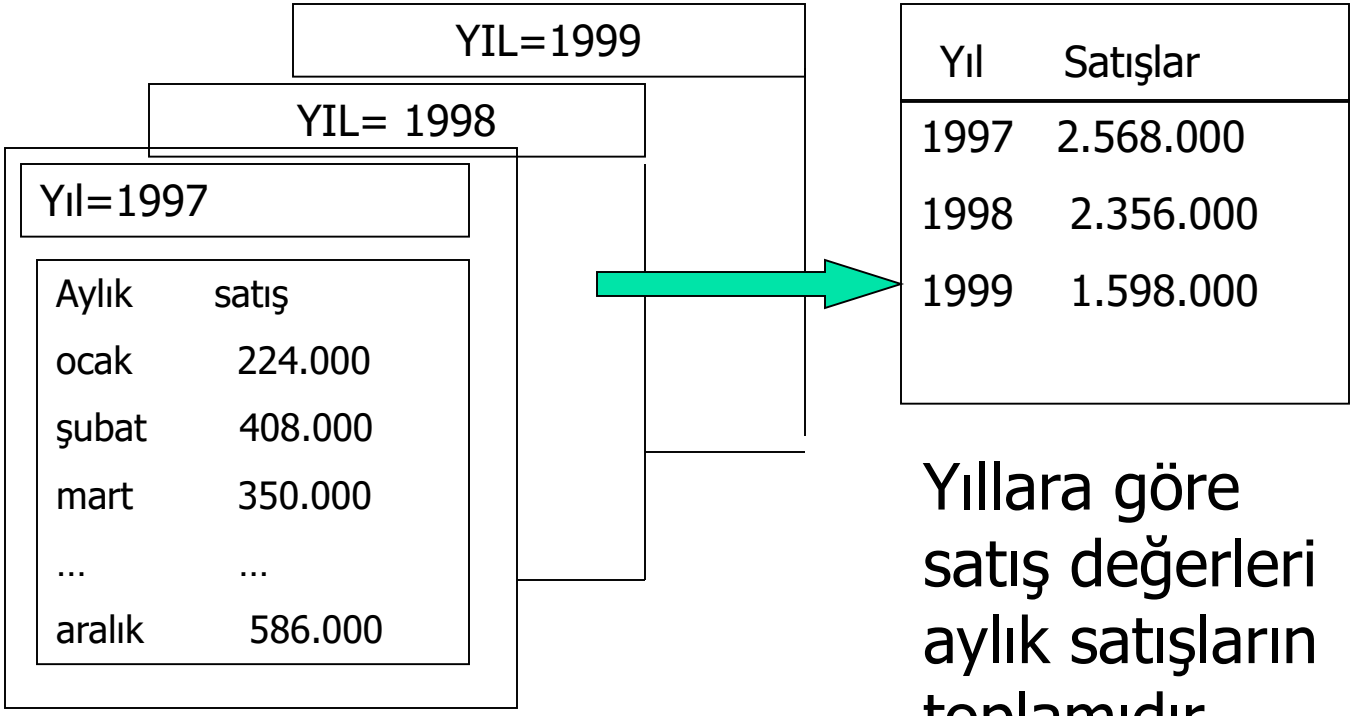
- Neden veri küçültme gerekiyor?
  - Veritabanı/veri ambarı çok büyük ola bilir
  - Büyük sayıda veriler üzerinde karmaşık analizler çok zaman gerektiriyor
- Veri küçültme
  - Aynı (veya hemen hemen aynı) analitik sonuçları veren , fakat daha küçük boyutlu veri kümesinin alınması
- Veri küçültme stratejileri
  - Veri küpünde toplama
  - Boyutsal küçültme — önemsiz özelliklerin silinmesi
  - Veri sıkıştırma
  - Ayrıklaştırma ve kavram hiyerarşisi

Veri Küpü Yığılması-Data Cube Aggregation

- Veri küpünün en aşağı seviyesi- temel küp (base cuboid)
  - İlgi alanı için verilerin bir yere yığılması
- Veri küplerinde çok seviyeli yığılma
  - Yukarı seviyelere doğru veri boyutu küçülüyor
- Uygun seviyeye erişim
  - Sorunun çözümü için yeterli olacak en küçük sunum seviyesini seçmeli

Verilerin özetlenmesi

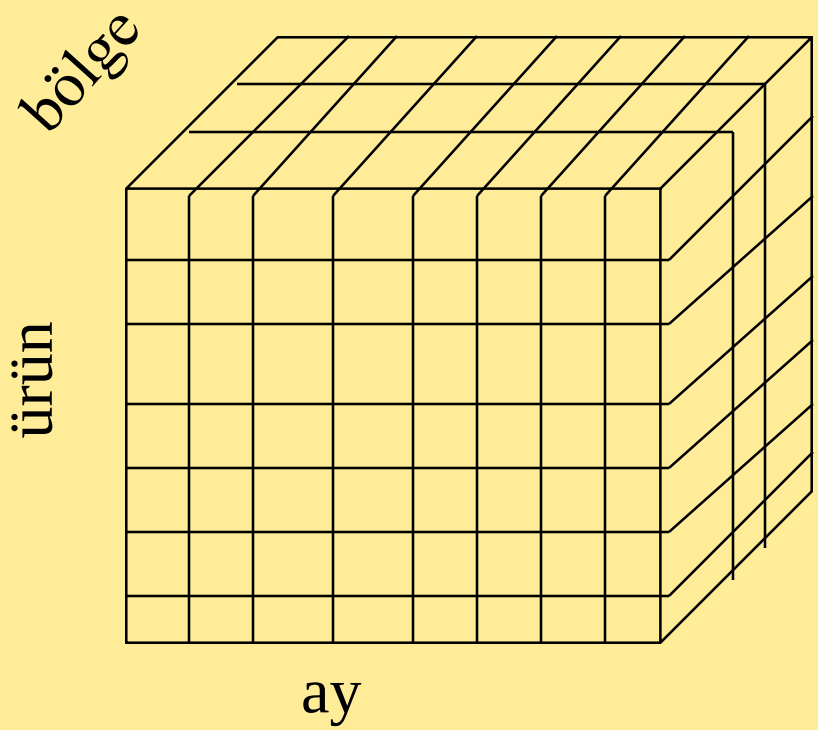
Üst yöneticilerin karar vermeleri için işletimsel (günlük ,aylık) veriler değil, özetlenmiş veriler daha önemlidir



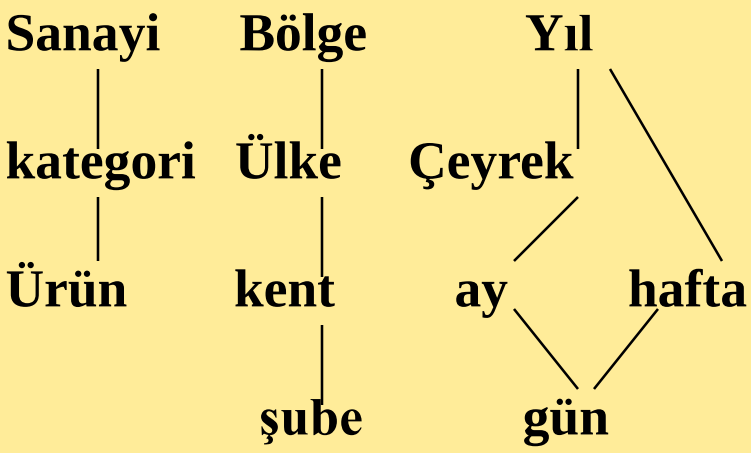
Yıllara göre satış değerleri aylık satışların toplamıdır

Çokboyutlu veriler

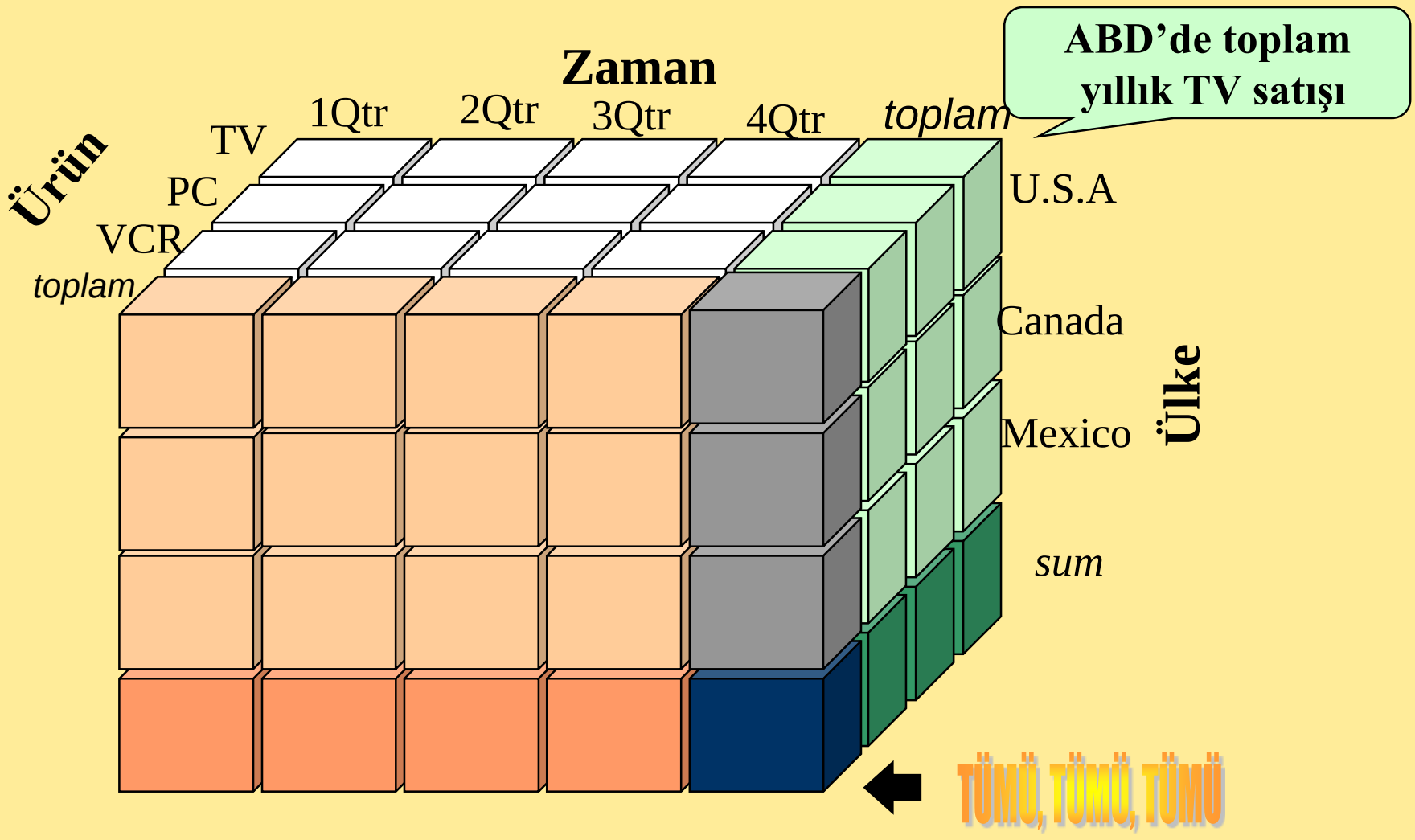
- Satış hacmi, ürün, ay ve bölge değerlerinin fonksiyonudur



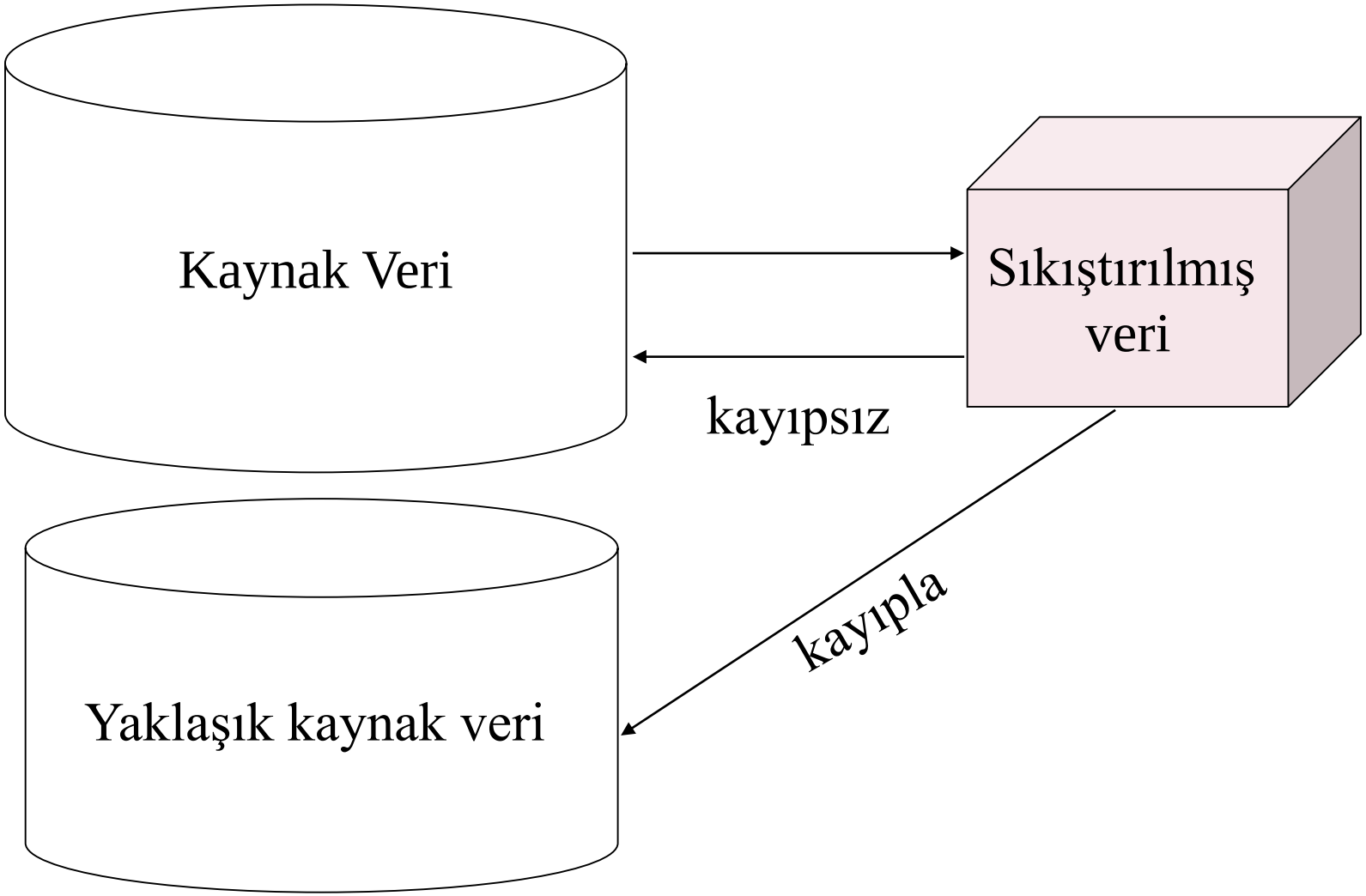
Boyutlar: Ürün, Mekan, Zaman
Yolların hiyerarşik özetlenmesi



Basit veri küpü



Veri sıkıştırma



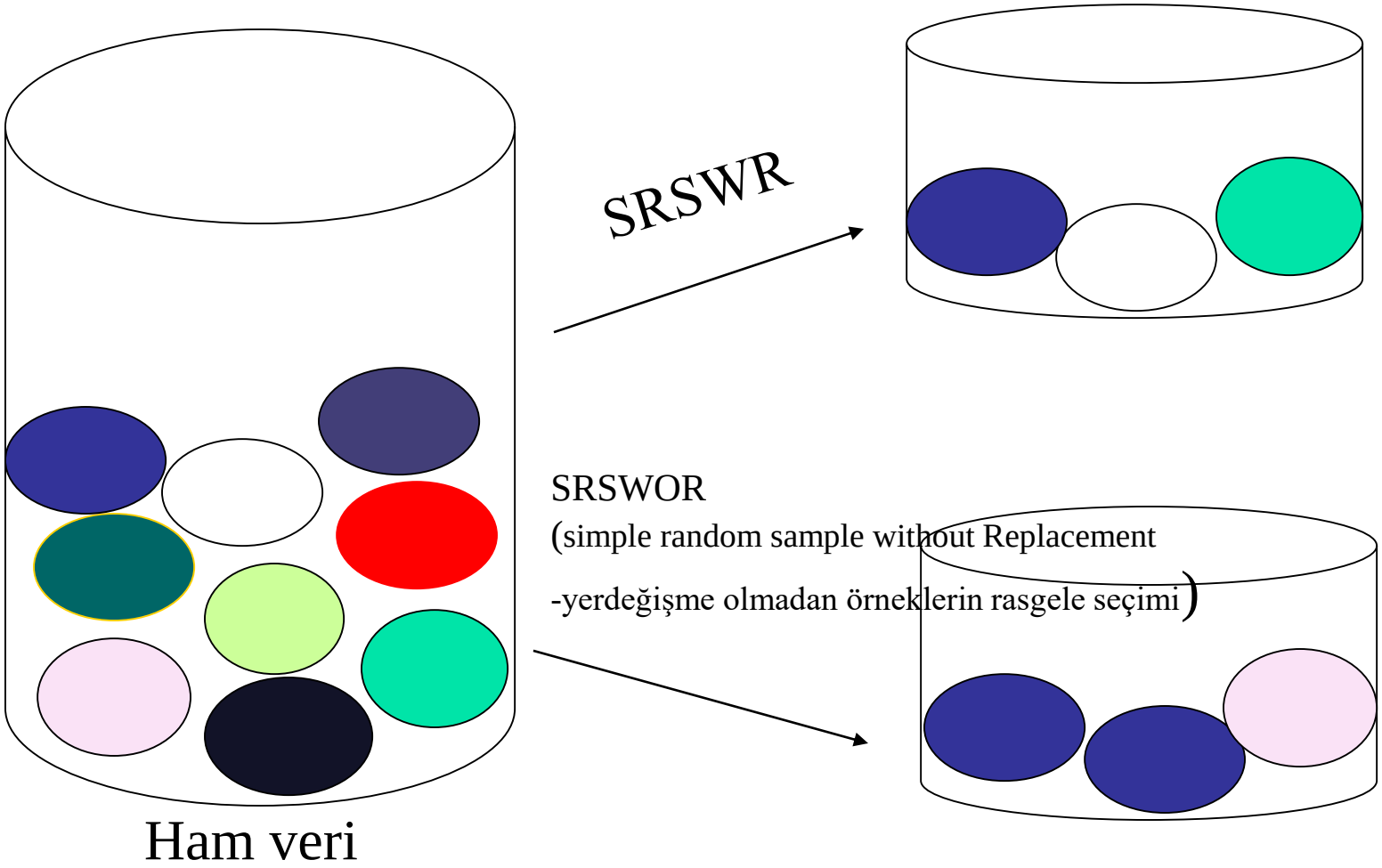
Veri küçültme yöntemi: Kümeleme

- Verilerin benzerliklerine göre kümelere dağıtılması
- Çokseviyeli kümeleme mümkündür; bu halde kümeler çok boyutlu ağaç yapıları indeksleri ile sunulur
- Çeşitli kümeleme algoritmaları mevcuttur

Veri küçültme Yöntemi: Örnekleme

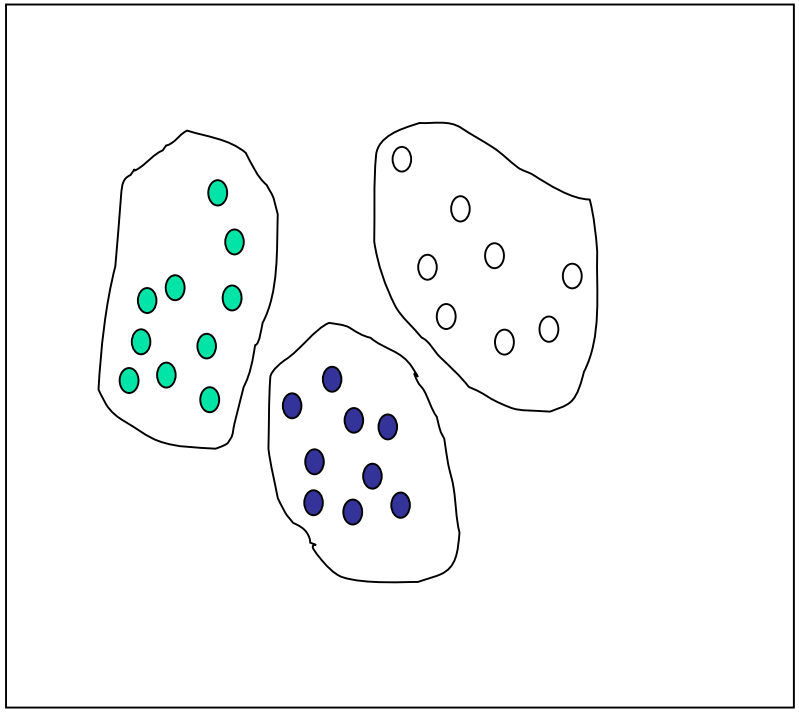
- Örnekleme: *N sayıda veriden oluşan* tam veri kümesini ifade etmek için küçük *s* örneğinin elde edilmesi
- Veri kümesini temsil edecek altkümenin seçilmesi
  - Basit rastgele seçim iyi sonuçlar vermeye bilir
  - Bütün veri tabanında kümelerin örneklerinin temsil oranlarını yakınlaştırmalı

Örnekleme: yerdeğişmeli ve yerdeğişmesiz

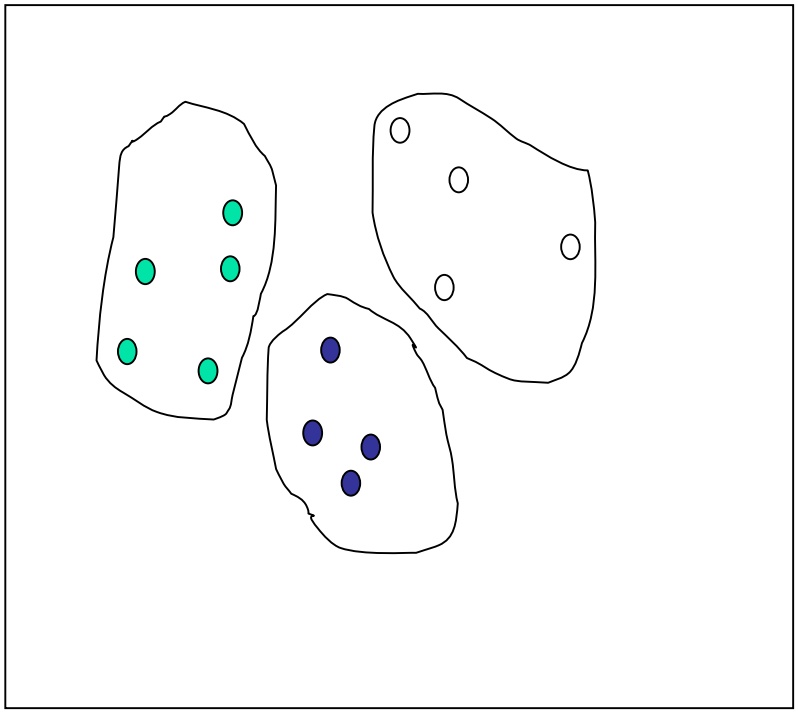


Örnekleme: Kümeleme

Ham veri



Küme



Yeni kümeler uygun kaynak kümelerden alınmış örneklerden oluşturulur

Verilerin Önişlenmesi

- Ayrıklaştırma ve kavram hiyerarşisi

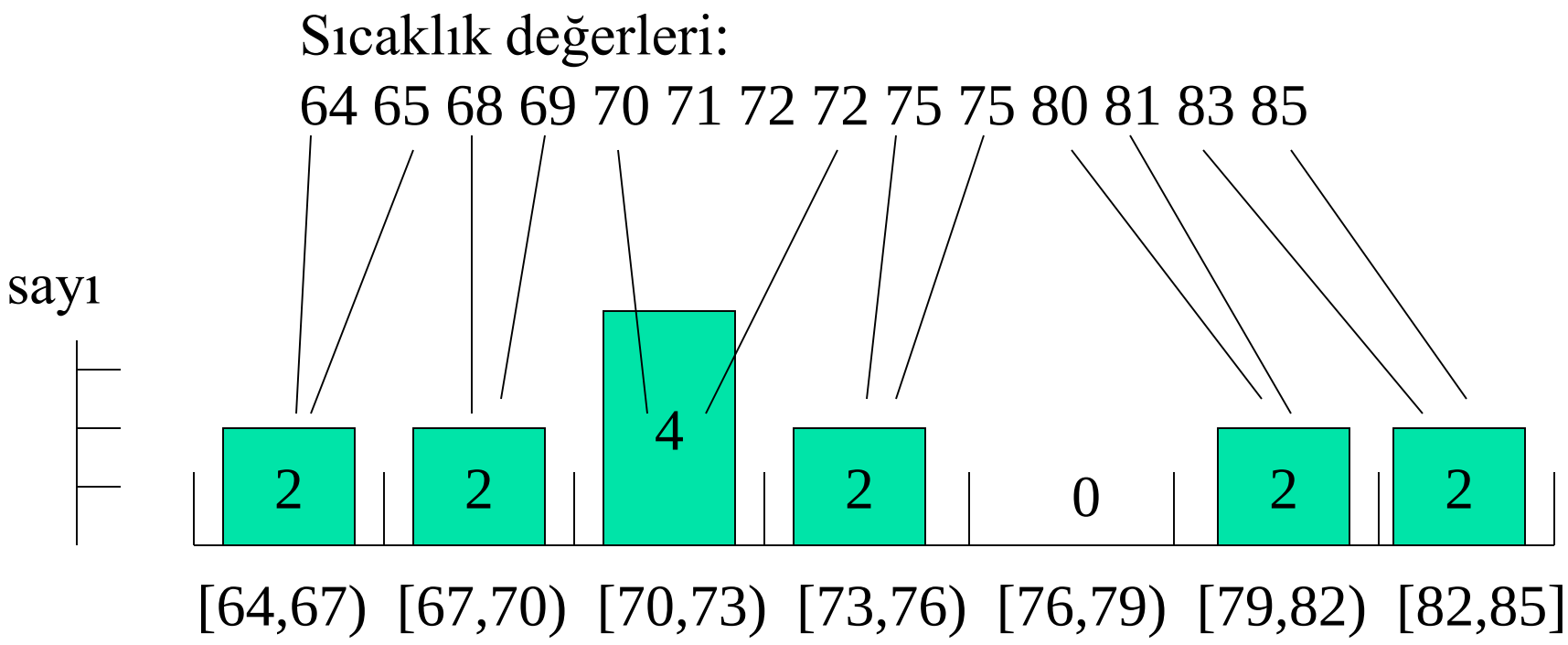
Ayrıklaştırma-Discretization

- Özelliklerin üç türü:
  - Nominal — sıralanmamış kümedeki değerler; örneğin, renk, meslek
  - Sıralı (Ordinal) — sıralanmış kümedeki değerler; örneğin, akademik unvanlar
  - Sürekli (Continuous) — gerçek sayılar;
- Ayrıklaştırma:
  - özelliklerin sürekli değer alanını aralıklara bölme
  - Ayrıklaştırma yolu ile verilerin boyutunu küçültme

Ayrıklaştırma ve kavram hiyerarşisi

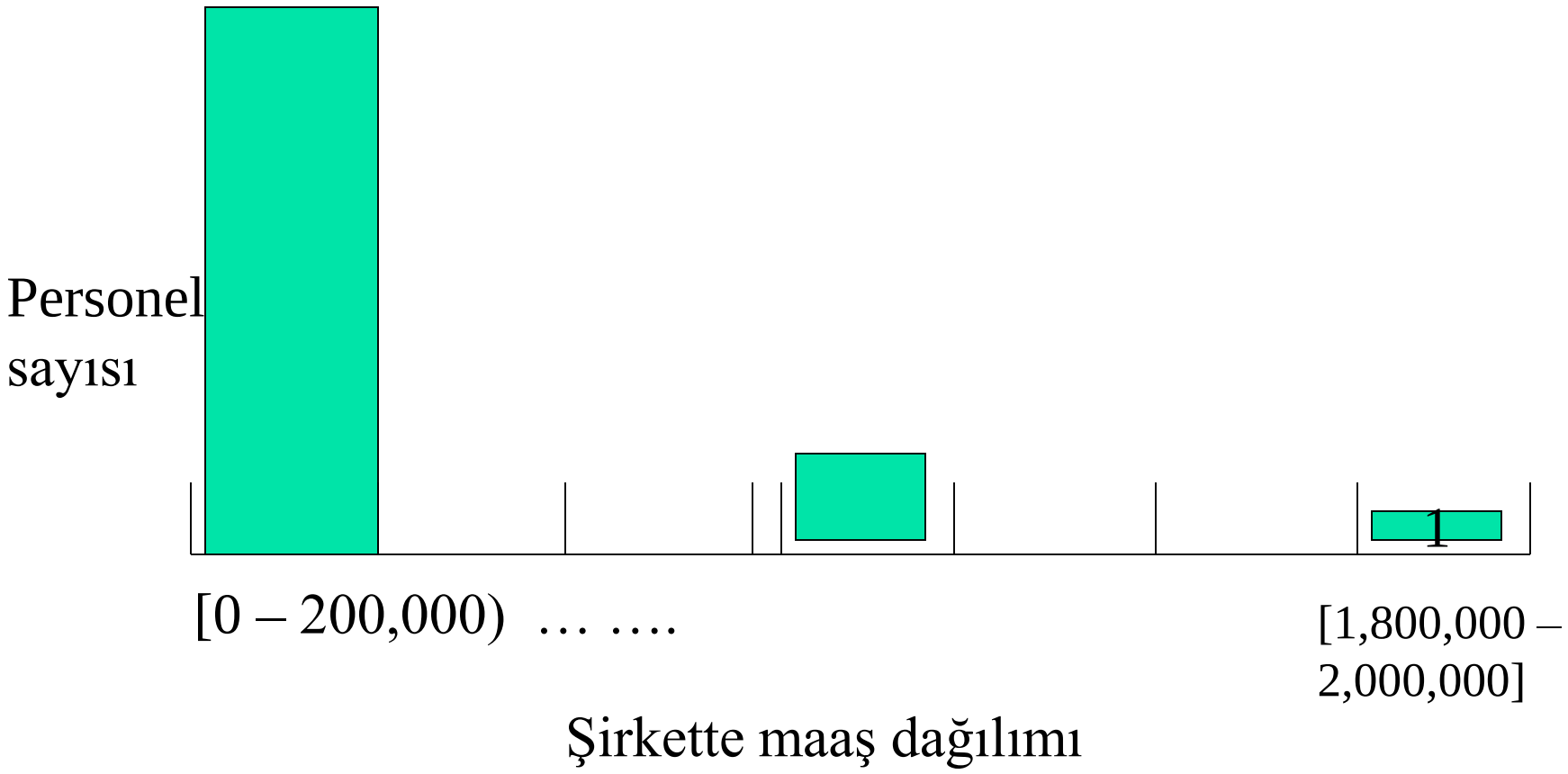
- ayrıklaştırma
  - Kesilmez türlü özelliğin değerler sayısını, değer alanını aralıklara bölmekle küçültmek
  - Aralık etiketleri (değerleri) gerçek veri değerlerinin yerine kullanıla bilir
  - Ayrıklaştırma , özelik üzerinde özyinelemeli olarak gerçekleştirile bilir
- Kavram hiyerarşisi
  - Aşağı seviye kavramlarını (örneğin, yaş için sayısal değerler)toplamak ve daha üst seviye kavramları ile (genç, orta yaşlı, yaşlı) değiştirmekle verilerin özyinelemeli olarak küçültülmesi

Ayrıklaştırma: Eşit genişlikli



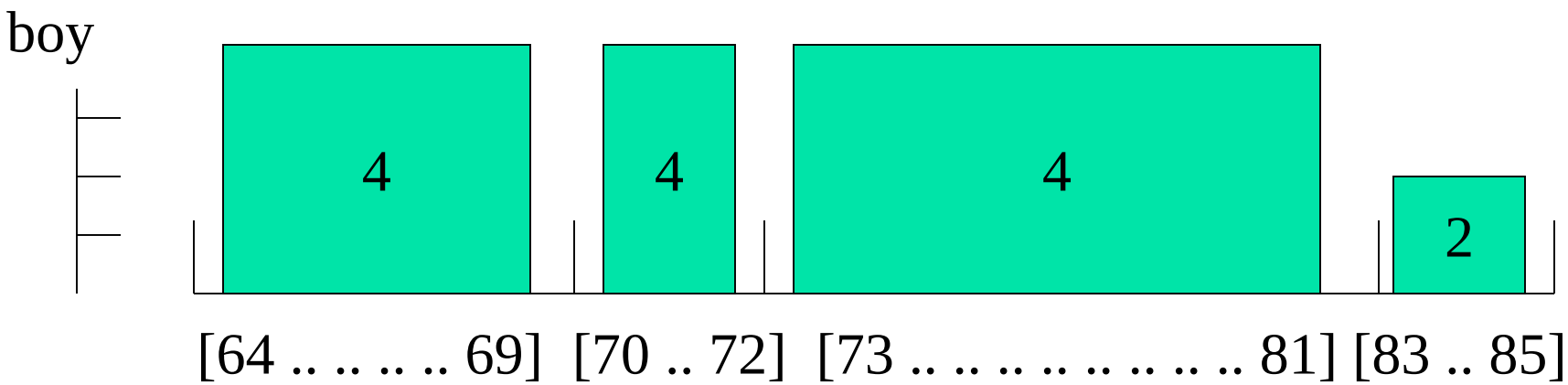
Çok sayıda veri yerine, bu verileri değerlerine göre eşit aralıklara bölmekle, veri dizininin aralıklarla ifade edilmesi

Eşit Genişlikli yöntem (2.örnek)



Eşit boylu

Sıcaklık değerleri:
64 65 68 69 70 71 72 72 75 75 80 81 83 85

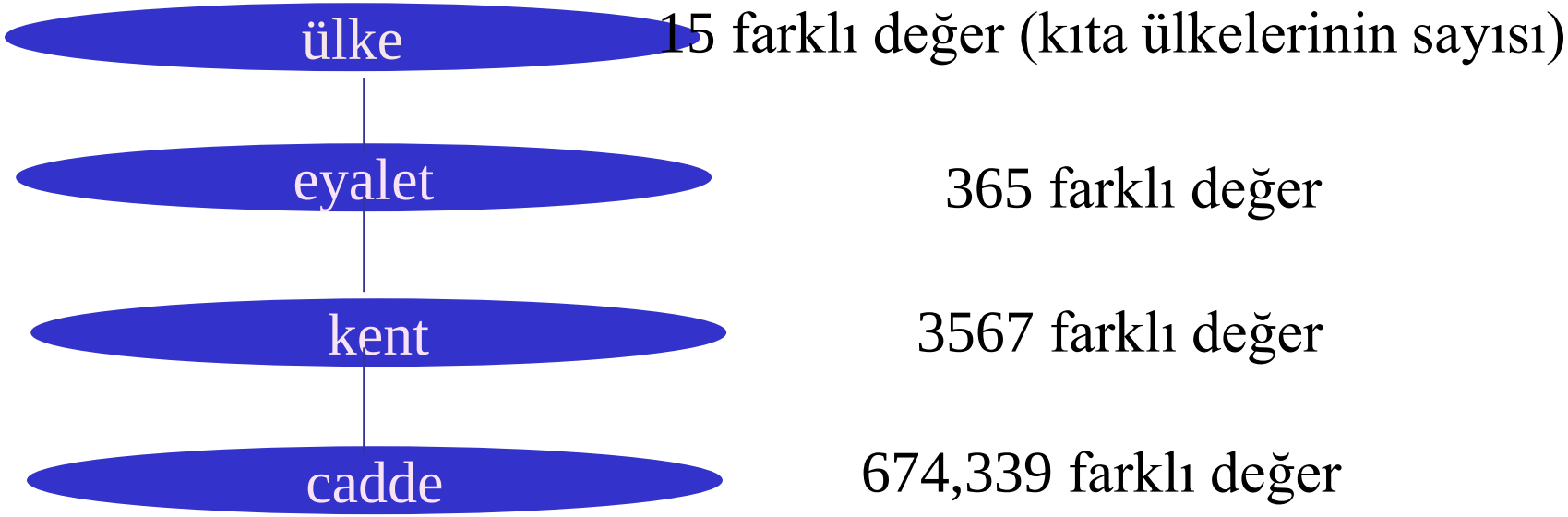


Boy= 4, yalnız sonuncu sepet dışında

Bu yöntemde tüm veri dizini eşit sayıda veri içeren aralıklarla ifade edilir

Kavram hiyerarşi

- Veri kümesinde her özellik üzere farklı değerler sayısını analiz etmekle hiyerarşileri üretmek mümkündür
  - En az farklı değeri bulunan özellik hiyerarşinin en üst seviyesine yerleştirilir
  - İstisnalar, örn., haftanın günleri, ay, çeyrek, yıl



İlaveler

- Zorunlu değildir, ama okunması gereklidir. Verilen örnekler sınavda yararlı olabilir

Verilerin niteliği neden düşüktür

Verilerin kalitesi çoğu zaman yüksek olmaz

- Neden**?**

– Veriler başkaları tarafından oluşturuluyor; sonra ise onları bütünleştirmek gerekiyor

– İnsanlar hata yapar

– İnsanlar çok meşgul olduklarından verilerin kalitesi onları çok düşündürmez («bu yeterlidir»)

Hata örnekleri

1,Dept. of Transportation, New York City,NY

2,Dept. of Finance,City of New York,NY

3,Office of Veteran's Affairs,New York,NY

- bu cümleleri tek biçimli ifade etmek gerekmektedir

Hata örnekleri

1,Dept. of Transportation,New York,NY
Two,Dept. of Finance,New York,NY
Office of Veteran's Affairs,
3,New York,NY

- hatalı numaralama

1,Dept. of Transportation,New York,NY

2,Dept. of Finance,New York,NY

3,Commission for the United Nations Consular Corps and Protocol,New York,NY

- 3. satırdaki cümle gerekenden fazla alan kapsamaktadır

Hata örnekleri

1,Dept. of Transportation,New York,NY

2,Dept. of Finance,New York,NY

2,Office of Veteran's Affairs,New York,NY

- Birincil anahtar (2) tekrarlanır

Biçimlendirme hataları

- zamanın farklı biçimlerde ifadesi:
  - 12/19/77
  - 12/19/1977
  - 12-19-77
  - 19/12/77
  - Dec 19, 1977
  - 19 December 1977
  - 9 in Tevet, 5738 (İbrani takvimi ile)

Farklı derecelendirme

– Bize gereken yaş aralığı 20-30, 30-40, 40-50, …

– Elimizdeki veriler ise : 15-30;0-45; 45-60,…aralığındadır

Veri Temizleme adımları

1. Yarım Yapılandırma
2. Standartlaştırma
3. Yerel tutarlılık yoklaması
4. Genel Tutarlılık yoklaması
5. Belge

Veri Temizleme adımlarına örnekler

Örnek «Kirli veriler»

Ralph Kimball *DBMS*, September 1996 kaynağından uyarlanmıştır

Yapısal olmayan dosyadan adres verileri:

Ralph B ve Julianne Kimball

Ste. 116

13150 Hiway 9

Box 1234 Boulder Crk

Colo 95006

Yarım-yapılandırma

Çözülme (parsing) de denir:

Addressee First Name(1): Ralph

Addressee Middle Initial(1): B

Addressee Last Name(1): Kimball

Addressee First Name(2): Julianne

Addressee Last Name(2): Kimball

Street Address Number: 13150

Street Name: Hiway 9

Suite Number: 116

Post Office Box Number: 1234

City: Boulder Crk

State: Colo

Five Digit Zip: 95006

Standartlaşma

- aynı anlamlı kelimeleri tek bir kelime ile ifade etmeliyiz

standard term

- Hiway 9 ‡ Highway 9
- Boulder Crk ‡ Boulder Creek
- Colo ‡ Colorado

Ralph B and
Julianne Kimball
Ste. 116
13150 Hiway 9
Box 1234 Boulder
Crk
Colo 95006

Yerel Tutarlılık yoklaması

Her veri parçası kendiliğinde bir anlam ifade ediyor mu?

- Boulder Creek ve Zip Code 95006 California eyaletindedir
- Devlet (State)

Colorado olarak gösterilmiştir

Ralph B and
Julianne Kimball
Ste. 116
13150 Hiway 9
Box 1234
Boulder Crk
Colo 95006

- 3 özellikten 2_si eyalet olarak California'nı gösteriyor.

Eyaleti (state) California olarak değişmeli

GENEL TUTARLILIK YOKLAMASI

- Ralph Kimball veya Julianne Kimball'ın kayıtlarını diğer müşteri kayıtlarında aramalı; adresteki tüm elementlerin aynı olduğuna eminlik sağlamalı
- Genel yoklamanın yerelden farkı, yalnız mevcut veri parçalarına değil, diğer parçalara da bakmasıdır

BELGELEME

Belge, metaverilerde yarım yapılandırma, standartlaştırma ve tutarlık yoklamaları yapma sonucudur

- Bütünleşik veritabanı oluşturmak için önemlidir
- Veritabanının gelecek güncellenmeleri için önemlidir