Necmettin Erbakan Üniversitesi
Bilgisayar Mühendisliği

Dr. Öğr. Üyesi Murat KARAKOYUN
(mkarakoyun@erbakan.edu.tr)

Ders: Veri Önişlemeye Giriş
[Konu: Aykırı Veri İşleme]

Sunu İçeriği

- Aykırı Veri Nedir?
- Normalizasyon ile İşleme
  - Min-Max Normalizasyonu
  - Z-Score Normalizasyonu
  - Ondalık Ölçekleme
- Veri Silme
  - IQR ile Silme
  - Chauvenet Kriterine Göre Silme

Aykırı Veri Nedir?

- Veri kümesinin ortalamasını ciddi anlamda etkileyen ve standart sapmanın yüksek çıkmasına sebep olan, verilerin genel yapısına uygun olmayan verilerdir.
- Gürültülü veri olarak da isimlendirilmektedir.

1. **Veri işlenerek kullanılmaya devam edilecek.**
   - 1.1. Min-Max Normalizasyonu
   - 1.2. Z-Score Normalizasyonu
   - 1.3. Ondalık Ölçekleme
2. **Veri silinecek.**
   - 2.1. IQR ile veri silinmesi
   - 2.2. Chauvenet ile veri silinmesi

1. Normalizasyon

Veri kümesindeki aykırı değerlerin ortalama ve standart sapmaya olan etkilerini azaltmak ve veri benzerliğini arttırmak amacıyla; verilerin belirli bir aralıkta tekrar düzenlenmesi işlemidir.

1.1. Min-Max Normalizasyonu

Veri kümesindeki değerleri doğrusal bir dönüşüm kullanarak belirlenen yeni sınırlar içerisine çeken normalizasyon yöntemidir.

Orijinal veri setinde: [minA, maxA]

Normalizasyon için belirlenen yeni aralık: [new_minA, new_maxA]

V: Orijinal veri setinde bir değer

V': Normalize edilmiş değer

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

1.1. Min-Max Normalizasyonu

Bir firmadaki maaş değerleri: 12000-98000 arasında olsun. Maaşlardan oluşan bu veri seti [0, 1] aralığında normalize edilmek istensin. Bu durumda 73600 olan bir maaşın normalize edilmiş değeri ne olur?

Orijinal veri setinde: [minA = 12000, maxA = 98000]

Normalizasyon için belirlenen yeni aralık: [new_minA = 0, new_maxA=1]

V: 73600

V': ?

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

1.2. Z-Score Normalizasyonu

Veri kümesindeki değerleri, ortalama ve standart sapma değerini kullanarak değiştiren normalizasyon yaklaşımıdır.

μ_A: Ortalama

σ_A: Standart sapma

v: Orijinal veri setinde bir değer

v': Normalize edilmiş değer

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

Ortalaması (μ) 54,000, standart sapması (σ) 16,000 olan bir veri setinde 73,600 değerinin Z-score' a göre normalize edilmiş değeri nedir?

$$\frac{73{,}600 - 54{,}000}{16{,}000} = 1.225$$

1.3. Ondalık Ölçekleme

Veri kümesindeki mutlak değerce en büyük veri kullanılarak verilerin [-1, 1] aralığına dönüştürüldüğü normalizasyon yöntemidir.

$$v' = \frac{v}{10^j}$$

j değeri Max(|v'|) < 1 şartını sağlatan en küçük tam sayıdır.

Örnek: Veri setindeki en küçük sayı -834 ve en büyük sayı 435 iken ondalık ölçeklemeyi yapacak **j** değeri nedir?

Max(|v|) =834 ⟶ Max(|v'|) < 1 şartını sağlamak için min(**j**) = **3**

-834 değeri -> v' = $-834/10^3$ = **-0.834**

2. Veri Silinmesi

Veri kümesindeki aykırı değerlerin veri setinden kaldırılarak değerlendirmeye alınmamasına dayalı yaklaşımlardır.

2.1. IQR Değeri ile Veri Silinmesi

- Bu yöntemde ilk çeyrek ve üçüncü çeyrek kullanılarak aykırı değerler tespit edilmektedir.

- Aykırı değerin tespiti: Q_3 değerinden IQR*1.5 miktardan daha büyük olan değerler veya Q_1 değerinden IQR*1.5 miktardan daha küçük değerlerdir.

2.1. IQR Değeri ile Veri Silinmesi

Veri kümesi: 1, 2, 2, 2, 3, 3, 4, 5, 5, 5, 6, 6, 6, 6, 7, 8, 8, 9, 27

Q_1 = (25/100)*(19+1) = **5**

Q_1 = 3

Q_3= (75/100)*(19+1)= **15**

Q_3 = 7

IQR = Q_3 - Q_1 = **4**

2.1. IQR Değeri ile Veri Silinmesi

Veri kümesi: 1, 2, 2, 2, 3, 3, 4, 5, 5, 5, 6, 6, 6, 6, 7, 8, 8, 9, 27

$Q_1 = 3$, $Q_3 = 7$, **IQR** = 4

Aykırı değerler için alt ve üst sınırları bulalım.

Alt sınır: $Q_1 - 1.5*IQR = 3 - 1.5*4 = -3$

Üst sınır: $Q_3 + 1.5*IQR = 7 + 1.5*4 = 13$

Sonuç: Veri kümesindeki **27** değeri aykırı bir değerdir.

2.2. Chauvenet ile Veri Silinmesi

- Az sayıda aykırı verinin bulunduğu veri setlerinde kullanılan bir yöntemdir.
- Chauvenet kriteri ortalama değerin her iki yanında 2σ aralığının dışında kalan ölçüm sonuçlarının elimine edilmesine gerek olup olmadığını belirlemede kullanılır.
- Her uygulama esnasında veri setindeki 1 veri için sonuç elde edilir.

2.2. Chauvenet ile Veri Silinmesi

- Her bir ölçüm için (maksimum sapma (d_{max})/standart sapma oranı (σ)) hesaplanmalıdır.

- Ölçülen değer Chauvenet kriterinden büyükse o ölçüm analiz dışı tutulmalıdır yani silinmelidir.

2.2. Chauvenet ile Veri Silinmesi

- Yapılan ölçüm sayısına bağlı olan Chauvenet kriterleri aşağıdaki tabloda verilmiştir.

Ölçüm sayısı	(Chauvenet Kriteri) d_{max}/σ
2	1.15
3	1.38
4	1.54
5	1.65
6	1.73
7	1.80
10	1.96
15	2.13
25	2.33
50	2.57
100	2.81
300	3.14
500	3.29
1000	3.48

2.2. Chauvenet ile Veri Silinmesi

Örnek: [5.30 5.73 6.77 5.26 4.33 5.45 6.09 5.64 5.81 5.75]

i	x	$\bar{x}$	$d = x - \bar{x}$	d^2	σ	2σ	d/σ
1	5.30	5.613	-0.313	0.009797	0.595	1.189	0.526
2	5.73	5.613	0.117	0.01369	0.595	1.189	0.197
3	6.77	5.613	1.157	1.33864	0.595	1.189	1.945
4	5.26	5.613	-0.353	0.12461	0.595	1.189	0.593
5	4.33	5.613	-1.283	1.64866	0.595	1.189	2.156
6	5.45	5.613	-0.163	0.02657	0.595	1.189	0.274
7	6.09	5.613	0.477	0.21753	0.595	1.189	0.802
8	5.64	5.613	0.027	0.000729	0.595	1.189	0.045
9	5.81	5.613	0.197	0.03881	0.595	1.189	0.331
10	5.75	5.613	0.137	0.01877	0.595	1.189	0.230
	Σ_1=56.13			**Σ_2=3.536**			

[d_{max} = 1.283 >? 2σ = 1.189] durumu kontrol edilmelidir.

[$d_{max} > 2\sigma$] olduğundan 5.örnek tablodan kontrol edilmelidir.

[d/σ = 2.156] değeri tablodaki kriter değeri **(1.96)** ile karşılaştırılmalıdır.

[d/σ = 2.156 > 1.96] olduğundan 5. örnek veri setinden çıkarılmalıdır.

2.2. Chauvenet ile Veri Silinmesi

- 5. örnek ihmal edilerek tekrar standart sapma hesaplanırsa 0.458 bulunur. Bu ilk değer olan 0.595 ile kıyaslanırsa değerin % 25 oranında değiştiği ve verilerin birbirine daha çok benzediği görülür.

- Bu değerden başka hatalı ölçümler olması mümkün olduğu için aynı işlem bir adım daha devam ettirilir hatalı başka nokta varsa çıkarılır yoksa işlem sonlandırılır.

Özet

Bu sunu kapsamında bir veri setindeki aykırı verilerin nasıl işleneceği ele alınmıştır.

Bu veriler normalizasyon yöntemleri ile tekrar elde edilerek kullanılabileceği gibi farklı yöntemler ile analiz edilerek veri setinden silinip silinmeyeceğine karar verilebilir.