VERİ MADENCİLİĞİ DERSİ

Dr.Öğr.Üyesi Mehmet Akif ŞAHMAN

Kümeleme Analizi

- Kümeleme analizi, sınıflandırmada olduğu gibi eldeki verilerin gruplara ayırması için kullanılır.
- Sınıflandırma işleminde sınıflar önceden belli iken kümeleme işleminde gruplar önceden belli değildir.
- Mevcut verilerin gruplara/kümelere, hatta kaç değişik gruba ayrılacağı eldeki verilerin benzerliğine göre belirlenir.
- Belirlenen her bir gruba küme ismi verilir.

Örn. Yaşları tutulan müşterilerin 20, 22, 26, 27, 40, 45, 46, 47, 49 olduğu düşünülürse, 20-27 yaşındakiler bir gruba, 40-49 yaşındakiler başka bir gruba dahil olurlar. Eğer eldeki müşteri verileri, 19, 20, 21, 21, 21, 26, 26, 26 ,27, 27 ,28 şeklinde olursa 19-21 yaşlarındaki bir gruba, 26-28 yaşındakiler diğer grupta toplanacaktır. İlk örnekte 20 ile 27 yaşındakiler aynı tutulurken diğer örnekte aynı kümede olmayacakları açıktır.

Kümeleme Analizi

- **Benzerlik ve Uzaklık :**

Veri tabanındaki veriler kümelere ayrılırken, benzerlik ve uzaklık kavramlarından faydalanılır. Benzerlik ve uzaklık bireysel olarak veriler için bakılabildiği gibi kümeler arasında da bakılabilir. Kümeleme işlemleri yapılırken benzer kümeler birleştirilebilir veya detaylandırmak için ayrılabilir, bu işlemin yapılabilmesi için gene kümelerin benzerlik ve uzaklık verilerinden faydalanılır. Mesafe aşağıdaki formül kullanılarak hesaplanır.

$$mes(X_m, X_j) = \sqrt{\sum_{i=1}^{n} (x_{mi} - x_{ji})^2}$$

Bu mesafeye **Euclid** mesafesi denir.

Kümeleme Analizi

- **Benzerlik ve Uzaklık :**

Benzerlik kavramı ise mesafenin tersi bir anlam içerir ve iki veri arasındaki yakınlığı gösterir. Genel olarak,

$$ben(X_m, X_j) = \frac{1}{1 + mes(X_m, X_j)}$$

Şeklide ifade edilir.

Kümeleme Analizi

- **Benzerlik ve Uzaklık :**

Benzerlik ölçümü için başka yöntemler de önerilmiştir. Bunlar,

$$ben(X_m, X_j)_{DICE} = \frac{2\sum_{i=1}^{n} x_{mi} x_{ji}}{\sum_{i=1}^{n} x_{mi}{}^2 + \sum_{i=1}^{n} x_{ji}{}^2}$$

$$ben(X_m, X_j)_{JACCARD} = \frac{\sum_{i=1}^{n} x_{mi} x_{ji}}{\sum_{i=1}^{n} x_{mi}{}^2 + \sum_{i=1}^{n} x_{ji}{}^2 - \sum_{i=1}^{n} x_{mi} x_{ji}}$$

Kümeleme Analizi

- **Benzerlik ve Uzaklık :**

$$ben(X_m, X_j)_{COSINE} = \frac{\sum_{i=1}^{n} x_{mi} x_{ji}}{\sqrt{\sum_{i=1}^{n} x_{mi}^{\,2} \sum_{i=1}^{n} x_{ji}^{\,2}}}$$

$$ben(X_m, X_j)_{OVERLAP} = \frac{\sum_{i=1}^{n} x_{mi} x_{ji}}{\min\left(\sum_{i=1}^{n} x_{mi}^{\,2}, \sum_{i=1}^{n} x_{ji}^{\,2} \right)}$$

Kümeleme Analizi

- **Benzerlik ve Uzaklık :**

Parametreler	X1	X2	X1*X2	X1^2	X2^2
1	1	0	0	1	0
2	3	2	6	9	4
3	2	2	4	4	4
4	4	4	16	16	16
5	5	4	20	25	16
6	7	5	35	49	25
	Toplam		81	104	65

Kümeleme Analizi

- **Örn. : X1 :{1,3,2,4,5,7} ve X2:{0,2,2,4,4,5} için benzerliği DICE,JACCARD,COSINE ve OVERLAP metotlarını kullanılarak hesaplayınız.**

$$ben(X_m, X_j)_{DICE} = \frac{2\sum_{i=1}^{n} x_{mi} x_{ji}}{\sum_{i=1}^{n} x_{mi}^{\ 2} + \sum_{i=1}^{n} x_{ji}^{\ 2}} = \frac{2.81}{104+65} = 0.958$$

$$ben(X_m, X_j)_{JACCARD} = \frac{\sum_{i=1}^{n} x_{mi} x_{ji}}{\sum_{i=1}^{n} x_{mi}^{\ 2} + \sum_{i=1}^{n} x_{ji}^{\ 2} - \sum_{i=1}^{n} x_{mi} x_{ji}} = \frac{81}{104+65-81} = 0.92$$

Kümeleme Analizi

- **Örn. : X1 :{1,3,2,4,5,7} ve X2:{0,2,2,4,4,5} için benzerliği DICE,JACCARD,COSINE ve OVERLAP metotlarını kullanılarak hesaplayınız.**

$$ben(X_m, X_j)_{COSINE} = \frac{\sum_{i=1}^{n} x_{mi} x_{ji}}{\sqrt{\sum_{i=1}^{n} x_{mi}^{\,2} \sum_{i=1}^{n} x_{ji}^{\,2}}} = \frac{81}{\sqrt{104.65}} = 0.985$$

$$ben(X_m, X_j)_{OVERLAP} = \frac{\sum_{i=1}^{n} x_{mi} x_{ji}}{\min\left(\sum_{i=1}^{n} x_{mi}^{\,2}, \sum_{i=1}^{n} x_{ji}^{\,2} \right)} = \frac{81}{\min(104, 65)} = 1.24$$

Kümeleme Analizi

- Eğer benzerlik için veriler nümerik değilse benzerlik için kullanılabilecek formüller aşağıdaki formüllerle hesaplanmalıdır.

$$ben(X_m, X_j)_{DICE} = \frac{2\,|\,x_m \cap x_j\,|}{|\,x_m\,| + |\,x_j\,|}$$

$$ben(X_m, X_j)_{JACCARD} = \frac{|\,x_m \cap x_j\,|}{|\,x_m\,| \cup |\,x_j\,|}$$

$$ben(X_m, X_j)_{COSINE} = \frac{|\,x_m \cap x_j\,|}{\sqrt{|\,x_m\,| . |\,x_j\,|}}$$

$$ben(X_m, X_j)_{OVERLAP} = \frac{|\,x_m \cap x_j\,|}{\min(|\,x_m\,|, |\,x_j\,|)}$$

Kümeleme Analizi

- **Örn. : Xm={elma,muz,armut,üzüm} , Xj={muz,peynir,süt,ekmek} kümelerinin benzerliklerini DICE,JACCARD,COSINE ve OVERLAP metotlarını kullanılarak hesaplayınız.**

Çözüm :

Küme	Elma	Muz	Armut	Üzüm	Peynir	Süt	Ekmek
Xm	1	1	1	1	0	0	0
Xj	0	1	0	0	1	1	1

$$ben(X_m, X_j)_{DICE} = \frac{2\,|\,x_m \cap x_j\,|}{|\,x_m\,| + |\,x_j\,|} = \frac{2.1}{4+4} = 0.25$$

$$ben(X_m, X_j)_{JACCARD} = \frac{|\,x_m \cap x_j\,|}{|\,x_m\,| \cup |\,x_j\,|} = \frac{1}{7} = 0.14$$

Kümeleme Analizi

- **Örn. : Xm={elma,muz,armut,üzüm} , Xj={muz,peynir,süt,ekmek} kümelerinin benzerliklerini DICE,JACCARD,COSINE ve OVERLAP metotlarını kullanılarak hesaplayınız.**

Çözüm :

Küme	Elma	Muz	Armut	Üzüm	Peynir	Süt	Ekmek
Xm	1	1	1	1	0	0	0
Xj	0	1	0	0	1	1	1

$$ben(X_m, X_j)_{COSINE} = \frac{|x_m \cap x_j|}{\sqrt{|x_m| . |x_j|}} = \frac{1}{\sqrt{4.4}} = 0.25$$

$$ben(X_m, X_j)_{OVERLAP} = \frac{|x_m \cap x_j|}{\min(|x_m|, |x_j|)} = \frac{1}{4} = 0.25$$

Kümeleme Analizi

- **K-Ortalama (K-Means) Algoritması:**

Bu algoritma sürekli olarak kümelerin yenilendiği ve en uygun çözüme ulaşılana kadar devam eden döngüsel bir algoritmadır. Algoritmanın pseudo (kaba) kodu aşağıdaki gibidir.

Girdiler

D={t1,t2,t3,…,tn} //eldeki veri tabanı

K // verilen küme sayısı

Algoritma

Keyfi olarak m1,m2,m3 ortalamalarını belirle,

Her bir t 'yi en yakın olduğu olduğu m kümesine ata

Kümelere ait ortalama değerini yeniden hesapla(m1,m2,m3…)

Kümelerde elemanlarında ortalama hesabından sonra bir değişiklik yoksa dur.

ilk adımdan itibaren tekrar et.

Çıktı

K adet küme

Kümeleme Analizi

- **Örn. : D={3,7,25,28,35,12,15,17,32,4} verilen verileri K-Means algoritmasını kullanarak iki kümeye atayınız.**

Öncelikle iki küme olacağı için rastgele iki tane m1=3 ve m2=7 ortalama değeri rastgele olarak seçilir. Bu ortalama değerlerine yakın olan veriler k1 ve k2 kümelerine aşağıdaki gibi atanır.

k1={3,4} , k2={7,25,28,35,12,15,17,32} daha sonra oluşan bu kümelerin yeni ortalamaları bulunur. m1=3.5 ve m2=21.3 olarak bulunur. Yeni kümeler k1 ve k2 aşağıdaki gibi değişir.

k1={3,4,7,12} , k2={25,28,35,15,17,32} yeni kümelerin ortalamaları m1=6.5 ve m2=25.3 olarak bulunur.

Kümeleme Analizi

- **Örn. : D={3,7,25,28,35,12,15,17,32,4} verilen verileri K-Means algoritmasını kullanarak iki kümeye atayınız.**

k1={3,4,7,12} , k2={25,28,35,15,17,32} yeni kümelerin ortalamaları m1=6.5 ve m2=25.3 olarak bulunur. Buna göre;

k1={3,4,7,12,15}, k2={25,28,35,17,32}

m1=8.2 ve m2=27.4 olarak bulunur.

k1={3,4,7,12,15,17} , k2={25,28,35,32} olarak kümeler belirlenir.

M1= 9.6 ve m2=30 olarak ortalamalar bulunur fakat bu yeni ortalama değerlerinden sonra kümelerde bir değişiklik olmadığı için algoritma sonlandırılır.