13. HAFTA

KÜMELEME ANALİZİ

Kümeleme Analizi de, Diskriminant analizi gibi bir sınıflandırma yöntemidir. Diskriminant analizinde birimlerin sınıflanacağı gruplar(kümeler) bilinmesine rağmen, kümeleme analizinde kümeler hakkında bir bilgi yoktur. Kümeleme analizinde birimler benzerliklerine veya farklılıklarına göre kümelere ayrılmaya çalışılır. Birimler 1'den, toplam birim sayı arasında kümeye ayrılmaya çalışılır. Öncelikle bazı benzerlik ölçülerini kısaca verelim.

Benzerlik Ölçüleri

Karmaşık veri yapılarından daha basit grup yapıları elde etmek için yakınlık veya benzerlik ölçülerine ihtiyaç duyulur. Birimlerin bir birine yakınlığı, uzaklık ölçülerine göre belirlenirken, değişkenlerin bir birine benzerliği ilişki katsayılarına göre belirlenebilmektedir.

İki tane p-boyutlu gözlem(birim) $\underline{\mathbf{x}}' = (x_1, x_2, ..., x_p)$ ve $\underline{\mathbf{y}}' = (y_1, y_2, ..., y_p)$ arasındaki oklid uzaklığı

$$d(\underline{\mathbf{x}}, \underline{\mathbf{y}}) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + ... + (x_p - y_p)^2}$$
$$= \sqrt{(\underline{\mathbf{x}} - \underline{\mathbf{y}})'(\underline{\mathbf{x}} - \underline{\mathbf{y}})}$$

biçiminde tanımlanmaktadır.

Ayni iki gözlem arasındaki istatistiksel uzaklık ise

$$d(\underline{\mathbf{x}}, \underline{\mathbf{y}}) = \sqrt{(\underline{\mathbf{x}} - \underline{\mathbf{y}})' \mathrm{A} (\underline{\mathbf{x}} - \underline{\mathbf{y}})}$$

olarak verilebilir. Burada $\mathrm{A} = \mathrm{S}^{-1}$ alınabilir. Ancak farklı gruplar hakkında önbilgi olmadığında, bu değer hesaplanamayacağından, öklid uzaklığı kümeleme için tercih edilir.

Diğer bir uzaklık ölçüsü

$$d(\underline{\mathbf{x}}, \underline{\mathbf{y}}) = \left[\sum_{i=1}^{p} |x_i - y_i|^m \right]^{1/m}$$

biçiminde verilen Minkowski uzaklığıdır. $m = 1$ için $d(\underline{\mathbf{x}}, \underline{\mathbf{y}})$ p-boyutlu koordinat sistemindeki iki nokta arasındaki city-block uzaklık ölçüsüdür. $m = 2$ alınırsa oklid uzaklığı elde edilir. Genel olarak m değeri değiştikçe, büyük ve küçük farklar için verilen ağırlık değişir.

Değişkenler aynı ağırlıkta ölçeklendirilmemiş olması durumunda kullanılan ağırlıklı öklit uzaklığı

$$d(\underline{\mathbf{x}},\underline{\mathbf{y}}) = \sqrt{w_1^2(x_1-y_1)^2 + w_2^2(x_2-y_2)^2 + \ldots + w_p^2(x_p-y_p)^2}$$
$$= \sqrt{\sum_{i=1}^{p} w_i^2(x_i-y_i)^2}$$

biçiminde tanımlanır. Burada w_i , i inci değişkenin standart sapma değeri s_i 'nin ya da aralık uzunluk(range) değerinin tersidir. $w_i = \frac{1}{s_i}$ alınırsa, elde edilen uzaklığa Karl-Pearson uzaklığı adı verilir.

Diğer bir uzaklık ölçüsü

$$d(\underline{\mathbf{x}},\underline{\mathbf{y}}) = D^2 = (\underline{\mathbf{x}}-\underline{\mathbf{y}})' S^{-1} (\underline{\mathbf{x}}-\underline{\mathbf{y}})$$

ya da

$$d(\underline{\mathbf{x}},\underline{\mathbf{y}}) = D^2 = (\overline{\underline{\mathbf{x}}}-\overline{\underline{\mathbf{y}}})' S^{-1} (\overline{\underline{\mathbf{x}}}-\overline{\underline{\mathbf{y}}})$$

olarak verilen Mahalanobis kare uzaklığıdır.

Bunlardan başka

$$T^2 = \frac{n_1 n_2}{n_1+n_2} (\overline{\underline{\mathbf{x}}}-\overline{\underline{\mathbf{y}}})' S^{-1} (\overline{\underline{\mathbf{x}}}-\overline{\underline{\mathbf{y}}})$$

ile verilen Hotelling T^2 uzaklığı,

$$d(\underline{\mathbf{x}},\underline{\mathbf{y}}) = \frac{\sum_{i=1}^{p} |x_i - y_i|}{\sum_{i=1}^{p} (x_i + y_i)}$$

biçiminde verilen Canberra uzaklığı ve.

$$d(\underline{\mathbf{x}},\underline{\mathbf{y}}) = 1 - \frac{2\sum_{i=1}^{p} \min(x_i, y_i)}{\sum_{i=1}^{p} (x_i + y_i)}$$

olarak verilen Czekanowski katsayısıdır.

Pratikte değişkenlerin hepsi nicel (aralıklı veya oransal) ölçme düzeyinde olması mümkün değildir. Bazı değişkenlerin nitel (sınıflayıcı veya sıralayıcı) ölçme düzeyinde olabilir. Böyle durumlarda yukarıda verilen formüller direkt kullanılamayacağından uzaklıklar

$$w_i = \begin{cases} 1 & \text{, nicel verler için} \\ \dfrac{1}{i \text{ inci değişkenin aralık uzaklığı}} & \text{, nitel verler için} \end{cases}$$

olmak üzere

$$d(\underline{\mathbf{x}}, \underline{\mathbf{y}}) = \frac{1}{p} \sum_{i=1}^{p} w_i \left| x_i - y_i \right|$$

ile bulunur.

Kümeleme analizinde kullanılacak uzaklıkların en azından aşağıdaki özellikleri sağlaması gerekir:

P ve Q iki nokta olmak üzere;

$$d(P,Q) = d(Q,P)$$
$$d(P,Q) > 0, \quad P \neq Q$$
$$d(P,Q) = 0, \quad P = Q$$
$$d(P,Q) \leq d(P,R) + d(R,Q)$$

olmalıdır.

Ancak bu koşulların sağlanmadığı durumlar için de kümeleme algoritmaları geliştirilmiştir.

Bir çok uygulamada birimlerdense, değişkenler gruplandırılabilir. Değişkenler için benzerlik ölçüleri korelasyon katsayıları ile ifade edilir. Bununla birlikte bazı kümeleme analiz uygulamalarında negatif korelasyonlar yerine mutlak değerleri alınır.

Değişkenler ikili (binary) olduğunda veriler çapraz tablolarla verilebilir. Bu durumda birimlerdense, değişkenler kategorilerde yer alır. Her değişken çifti için tabloda kategorileştirilmiş n birim vardır. Bunlar 0 ve 1 ile kodlanmıştır. Çapraz tablo

		Değişken k		
		1	0	Toplam
Değişken i	1	a	b	$a+b$
	0	c	d	$c+d$
	Toplam	$a+c$	$b+d$	$n=a+b+c+d$

biçimindedir. Burada n birimden b tanesi, i 'nin 1'e ve k'nın 0'a eşit olanların sayısıdır. Çapraz tablodaki ikili değişkenlere ilişkin çarpımsal moment korelasyon değeri,

$$r = \frac{ad - bc}{\left[(a+b)(c+d)(a+c)(b+d)\right]^{1/2}}$$

ifadesine göre elde edilir. Bu eşitlikten elde edilecek değer iki değişken arasındaki benzerlik ölçüsü olarak alınabilir. Bu korelasyon katsayısı, iki kategorik değişkenin bağımsızlık testi için elde edilen Ki-kare test istatistiği $r^2 = \frac{\chi^2}{n}$ ile ilişkilidir.

Kümeleme Yöntemleri

Hiyerarşik ve Hiyerarşik olmayan kümeleme yöntemleri üzerinde durulacak.

Hiyerarşik (Aşama Sıralı) Kümeleme Yöntemleri

Hiyerarşik kümeleme tekniklerinde ya birbirini izleyen birleşme serileri ya da birbirini izleyen bölünme serileriyle işe başlanılır.

Toplamalı (Agglomeratve) hiyerarşik metotlar bireysel birimlerle işe başlar. Toplamalı hiyerarşik yöntemde, başlangıçta birim sayısı kadar bireysel küme vardır. Bir birine en çok benzeyen iki birim birleştirilerek, ilk grup (küme) oluşturulur. Benzerlikler azaldığında bütün alt gruplar tek bir kümede birleşir.

Bölen (Divisive) hiyerarşik yöntemler farklı yönde çalışır. Başlangıçta tüm birimler tek bir kümededir. Bu tek grup her birindeki birimler birbirinden oldukça uzak olacak şekilde iki alt gruba bölünür. Daha sonra bu alt gruplar birbirine benzemeyen alt gruplara bölünerek işlem

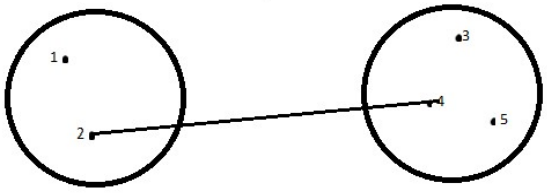
devam eder. Sonuçta her bir birim bir grubu (kümeyi) oluşturacak şekilde, birim sayısı kadar küme elde edilmiş olur.

Toplamalı ve Bölen yöntemlerinin sonuçları Dendogram olarak bilinen iki boyutlu diyagramda gösterilir. Dendogram, birimlerin birbirini izleyen seviyelerdeki birleşme ve bölünmelerini gösterir.

Bu derste, toplamalı hiyerarşik yöntemlerden bağlantı (Linkage) yöntemleri üzerinde durulacak.

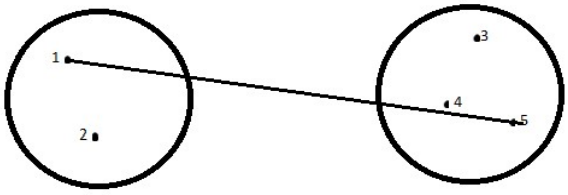
Bağlantı yöntemleri, değişkenlerde olduğu gibi küme birimleri için de uygundur. Ancak bu durum toplamalı hiyerarşik yöntemlerin hepsi için geçerli değildir. Bağlantı yöntemleri; tek bağlantı (single linkage), tam bağlantı (complete linkage) ve ortalama (orta) bağlantı (average linkage) olarak sınıflandırılabilir. Tek bağlantı yöntemi; en küçük uzaklık (minimum distance), en yakın komşuluk (nearest neighbor), Tam bağlantı; en büyük uzaklık (maximum distance), en uzak komşuluk (farthest neighbor) ve Ortalama bağlantı; ortalama uzaklık (average distance) olarak da adlandırılırlar.

Küme Uzaklığı



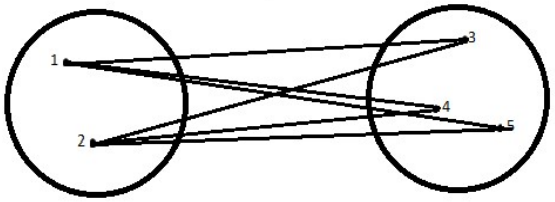
d_{24}

a. Tek Bağlantı



d_{15}

b. Tam Bağlantı



$$\frac{d_{13}+d_{14}+d_{15}+d_{23}+d_{24}+d_{25}}{6}$$

c. Ortalama Bağlantı

Tek bağlantıda gruplardaki birimlerden birbirine en yakın yani aralarındaki uzaklık en az olanların birleştirilmesidir. Tam bağlantıda gruplardaki birimlerden birbirine en uzak yani aralarındaki uzaklık en fazla olanların birleştirilmesidir. Ortalama bağlantıda ise, her gruptaki her birimin, diğer gruptaki birimler arasındaki uzaklıkların ortalamasına göre birleştirilirler.

Aşağıdaki algoritma, N tane birimin gruplandırılması için toplamalı hiyerarşik kümeleme adımlarını vermektedir:

1. Her biri tek bir birim içeren ve uzaklıkların (veya benzerliklerin) NxN simetrik matrisi $D = \{d_{ik}\}$ olan N küme ile işe başla.
2. Kümelerin en yakın (en çok benzer) çiftleri için uzaklık matrisini irdele. En çok benzer U ve V kümeleri arasındaki uzaklık d_{UV} olsun.
3. U ve V kümelerini birleştir. Elde edilen yeni kümeyi UV ile göster. Uzaklık matrisindeki elemanları
   a. U ve V kümelerine ilişkin satır ve sütunların çıkarılmasıyla,
   b. UV kümesi ile geriye kalan kümeler arasındaki uzaklıklarla verilen satır veya sütunların eklenmesiyle

   elde edilir.
4. Toplamda N-1 defa 2. ve 3. adımları tekrarla.

Algoritma bitiminde tüm birimler tek bir kümede birleşmiş olacaktır. Küme birimlerinin hangi uzaklık değerinde birleştiklerinin kayıt edilmesi gerekir.