14. HAFTA

Tek Bağlantı

Tek bağlantı algoritması için girdiler birim çiftleri arasındaki uzaklıklar veya benzerlikler olabilir. Gruplar en küçük uzaklık veya en büyük benzerliklere sahip birimlerin birleştirilmesiyle oluşturulurlar. İlk olarak, $D=\{d_{ik}\}$ matrisindeki en küçük uzaklık bulunur ve U ve V gibi ilişkili birimlerin (kümelerin) birleştirilmesiyle (UV) kümesi elde edilir. Daha sonra yeni oluşan (UV) kümesinin, diğer her hangi bir küme (birim) W ile arasındaki uzaklık

$$d_{(UV)W}=\min\{d_{UW},d_{VW}\}$$

ile hesaplanır. Buradaki d_{UW} ve d_{VW} değerleri U kümesini W kümesine ve V kümesinin W kümesine olan uzaklıklardır.

Tek bağlantı kümelemenin sonuçları dendogram veya ağaç diyagramı denilen grafikte gösterilir. Grafikteki kollar kümeleri göstermektedir.

Örnek 30: $n=5$ birimin aralarındaki ikişerli uzaklıklara göre elde edilen D uzaklık matrisi aşağıda verilmiştir. Bu beş birimi Tek Bağlantı kümeleme yöntemine göre kümeleyiniz.

$$D=\{d_{ik}\}=\begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array}\begin{array}{c} \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array} \\ \begin{bmatrix} 0 & & & & \\ 9 & 0 & & & \\ 3 & 7 & 0 & & \\ 6 & 5 & 9 & 0 & \\ 11 & 10 & 2 & 8 & 0 \end{bmatrix} \end{array}$$

Çözüm 30:

1. Öncelikle her birim bir küme olarak alınır. Aralarındaki uzaklık en az olan iki birim (küme) birleştirilir.

Uzaklık matrisinin elemanları incelendiğinde en küçük değer 2 dir. $\min\{d_{ik}\}=d_{35}=2$ olduğundan 5. ve 3. birimler birleştirilerek, (35) kümesi elde edilir.

Elde edilen yeni kümenin diğer kümelere uzaklıkları hesaplanır:

(35) kümesinin diğer kümelere uzaklıkları

$$d_{(35)1} = \min\{d_{31}, d_{51}\} = \min\{3,11\} = 3$$
$$d_{(35)2} = \min\{d_{32}, d_{52}\} = \min\{7,10\} = 7$$
$$d_{(35)4} = \min\{d_{34}, d_{54}\} = \min\{9,8\} = 8$$

biçimindedir.

D matrisinden 3. ve 5. birimlere ilişkin satır ve sütunları çıkarılıp, (35) kümesi eklenir. Böylece uzaklık matrisinde 1, 2, 4 ve (35) olmak üzere 4 küme vardır. (35) kümesini eklenmesiyle elde edilen yeni uzaklık matrisi

$$D = \{d_{ik}\} = \begin{array}{c} \\ (35) \\ 1 \\ 2 \\ 4 \end{array} \begin{array}{c} \begin{array}{cccc} (35) & 1 & 2 & 4 \end{array} \\ \begin{bmatrix} 0 & & & \\ 3 & 0 & & \\ 7 & 9 & 0 & \\ 8 & 6 & 5 & 0 \end{bmatrix} \end{array}$$

elde edilir.

2. En son elde edilen uzaklık matrisinin en küçük elemanı 3 dür. Yani

$$\min\{d_{ik}\} = d_{(35)1} = 3$$

olduğundan (35) ile 1. birimler (135) kümesi olarak birleştirilir.

Elde edilen yeni kümenin diğer kümelere uzaklıkları hesaplanır:

(135) kümesinin diğer kümelere uzaklıkları

$$d_{(135)2} = \min\{d_{12}, d_{(35)2}\} = \min\{9,7\} = 7$$
$$d_{(135)4} = \min\{d_{14}, d_{(35)4}\} = \min\{6,8\} = 6$$

biçimindedir.

D matrisinden (35). ve 1. birimlere ilişkin satır ve sütunları çıkarılıp, (135) kümesi eklenir. Böylece uzaklık matrisinde 2, 4 ve (135) olmak üzere 3 küme vardır. (135) kümesini eklenmesiyle elde edilen yeni uzaklık matrisi

$$D = \{d_{ik}\} = \begin{array}{c} \\ (135) \\ 2 \\ 4 \end{array} \begin{array}{c} \begin{array}{ccc} (135) & 2 & 4 \end{array} \\ \begin{bmatrix} 0 & & \\ 7 & 0 & \\ 6 & 5 & 0 \end{bmatrix} \end{array}$$

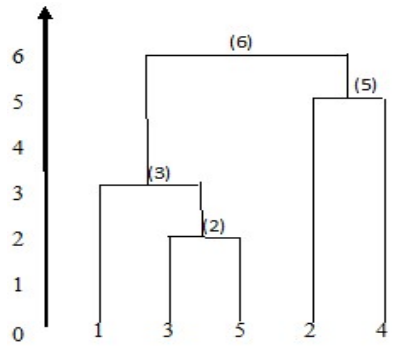
elde edilir.

3. En son elde edilen uzaklık matrisinin en küçük elemanı 5 dir. Yani $\min\{d_{ik}\} = d_{24} = 5$ olduğundan 2. ile 4. birimler (24) kümesi olarak birleştirilir. Bu noktada (135) ve (24) ile gösterilen iki farklı küme elde edilmiş olur. Bu kümelerin en yakın komşuluk uzaklığı $d_{(135)(24)} = \min\{d_{(135)2}, d_{(135)4}\} = \min\{7,6\} = 6$ dır. En son elde edilen uzaklık matrisi

$$D = \{d_{ik}\} = \begin{array}{c} \\ (135) \\ (24) \end{array} \begin{array}{c} \begin{array}{cc} (135) & (24) \end{array} \\ \begin{bmatrix} 0 & \\ 6 & 0 \end{bmatrix} \end{array}$$

4. Sonuç olarak (135) ve (24) kümeleri en yakın komşuluk uzaklığı 6 'ya ulaştığında (12345) tek bir küme biçiminde birleştirilir.

Dendogram:



Dendogram incelendiğinde, birimlerin 2 veya 3 kümeye ayrılabileceği söylenebilir.

Tam Bağlantı

Tam bağlantı kümeleme yöntemi, tek bağlantı yöntemiyle hemen hemen aynı mantıkla işlem çalışır. Ancak tek bağlantı yönteminden faklı olarak, tam bağlantı yönteminde bir kümenin, birleştirilmiş bir kümeye olan uzaklığı; bu kümenin en son birleştirilen kümelere(birimlere) olan uzaklık değerlerinin maksimumuna göre belirlenir. Her adımda kümeler arasındaki uzaklık(veya benzerlik), iki birim arasındaki uzaklık(veya benzerlik) ile belirlenir. Böylece tam bağlantı, bir kümedeki bütün birimlerin birbirleriyle olan maksimum uzaklık(veya minimum benzerlik) değerini verir.

Genel algoritma tek bağlantı yönteminde olduğu gibi $D = \{d_{ik}\}$ matrisindeki en küçük değerin bulunmasıyla başlar ve bu en küçük değere sahip olan U ve V gibi ilişkili birimlerin (kümelerin)

birleştirilmesiyle (UV) kümesi elde edilir. Yukarıda verilen algoritmanın 3. Adımı için (UV) kümesi ile diğer herhangi bir W kümesi arasındaki uzaklık

$$d_{(UV)W} = \max\{d_{UW}, d_{VW}\}$$

ile hesaplanır. Burada d_{UW} ve d_{VW} değerleri, U kümesini W kümesine ve V kümesinin W kümesine olan uzaklıklardır.

Örnek 31: $n = 5$ birimin aralarındaki ikişerli uzaklıklara göre elde edilen D uzaklık matrisi aşağıda verilmiştir. Bu beş birimi Tam Bağlantı kümeleme yöntemine göre kümeleyiniz.

$$D = \{d_{ik}\} = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array} \\ \begin{bmatrix} 0 & & & & \\ 9 & 0 & & & \\ 3 & 7 & 0 & & \\ 6 & 5 & 9 & 0 & \\ 11 & 10 & 2 & 8 & 0 \end{bmatrix} \end{array}$$

Çözüm 31:

1. Öncelikle her birim bir küme olarak alınır. Aralarındaki uzaklık en az olan iki birim (küme) birleştirilir.

Uzaklık matrisinin elemanları incelendiğinde en küçük değer 2 dir. $\min\{d_{ik}\} = d_{35} = 2$ olduğundan 5. ve 3. birimler birleştirilerek, (35) kümesi elde edilir.

Elde edilen yeni kümenin diğer kümelere uzaklıkları hesaplanır:

(35) kümesinin diğer kümelere uzaklıkları

$$d_{(35)1} = \max\{d_{31}, d_{51}\} = \max\{3,11\} = 11$$
$$d_{(35)2} = \max\{d_{32}, d_{52}\} = \max\{7,10\} = 10$$
$$d_{(35)4} = \max\{d_{34}, d_{54}\} = \max\{9,8\} = 9$$

biçimindedir.

D matrisinden 3. ve 5. birimlere ilişkin satır ve sütunları çıkarılıp, (35) kümesi eklenir. Böylece uzaklık matrisinde 1, 2, 4 ve (35) olmak üzere 4 küme vardır. (35) kümesini eklenmesiyle elde edilen yeni uzaklık matrisi

$$D=\{d_{ik}\}=\begin{array}{c} \\ (35) \\ 1 \\ 2 \\ 4 \end{array}\begin{array}{c} \begin{array}{cccc} (35) & 1 & 2 & 4 \end{array} \\ \begin{bmatrix} 0 & & & \\ 11 & 0 & & \\ 10 & 9 & 0 & \\ 9 & 6 & 5 & 0 \end{bmatrix} \end{array}$$

elde edilir.

2. En son elde edilen uzaklık matrisinin en küçük elemanı 5 dir. Yani $\min\{d_{ik}\}=d_{24}=5$ olduğundan 2. ile 4. birimler (24) kümesi olarak birleştirilir.

Elde edilen yeni kümenin diğer kümelere uzaklıkları hesaplanır:

(24) kümesinin diğer kümelere uzaklıkları

$$d_{(24)(35)}=\max\{d_{2(35)},d_{4(35)}\}=\max\{10,9\}=10$$

$$d_{(24)1}=\max\{d_{21},d_{41}\}=\max\{9,6\}=9$$

biçimindedir.

D matrisinden 2. ve 4. birimlere ilişkin satır ve sütunları çıkarılıp, (24) kümesi eklenir. Böylece uzaklık matrisinde (35), (24) ve 1 olmak üzere 3 küme vardır. (24) kümesinin eklenmesiyle elde edilen yeni uzaklık matrisi

$$D=\{d_{ik}\}=\begin{array}{c} \\ (35) \\ (24) \\ 1 \end{array}\begin{array}{c} \begin{array}{ccc} (35) & (24) & 1 \end{array} \\ \begin{bmatrix} 0 & & \\ 10 & 0 & \\ 11 & 9 & 0 \end{bmatrix} \end{array}$$

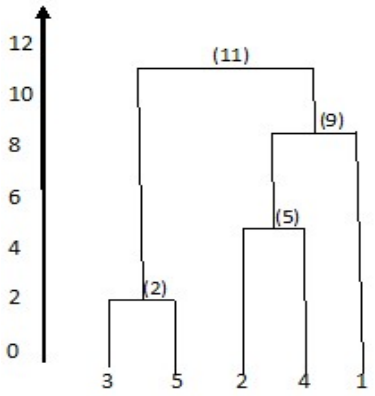
elde edilir.

3. En son elde edilen uzaklık matrisinin en küçük elemanı 9 dur. Yani $\min\{d_{ik}\}=d_{1(24)}=9$ olduğundan (24) kümesi ile 1. Küme birleştirilerek, (124) kümesi oluşturulur. Bu noktada (124) ve (35) ile gösterilen iki farklı küme elde edilmiş olur. Bu kümelerin en yakın komşuluk uzaklığı $d_{(124)(35)}=\max\{d_{1(35)},d_{(24)(35)}\}=\max\{11,10\}=11$ dir. En son elde edilen uzaklık matrisi

$$D=\{d_{ik}\}=\begin{array}{c} \\ (124) \\ (35) \end{array}\begin{array}{c} \begin{array}{cc} (124) & (35) \end{array} \\ \begin{bmatrix} 0 & \\ 11 & 0 \end{bmatrix} \end{array}$$

elde edilir. Sonuç olarak (124) ve (35) kümeleri en yakın komşuluk uzaklığı 11 ‘e ulaştığında (12345) tek bir küme biçiminde birleştirilir.

Dendogram:



Dendogram incelendiğinde, birimlerin 2 veya 3 kümeye ayrılabileceği söylenebilir.

Hiyerarşik Olmayan Kümeleme Yöntemleri

Hiyerarşik Olmayan Kümeleme Yönteminde, küme sayısı hakkında ön bilgi vardır veya araştırmacı anlamlı olacak bir küme sayısına karar vermiştir. Hiyerarşik olmayan kümeleme, hiyerarşik kümelemeye göre daha az zaman gerektirir. Hiyerarşik olmayan yöntemler, hiyerarşik yöntemlere göre daha büyük veri setlerine uygulanır.

Hiyerarşik olmayan yöntemler ya gruplardaki (kümelerdeki) birimlerin ilk parçalarından ya da kümelerin çekirdeklerinden oluşan çekirdek noktaların ilk kümesiyle işe başlar. İlk grupların(kümelerin) oluşturulması tamamen keyfidir. İlk kümelerin oluşturulmasındaki yollardan biri, birimler arasındaki çekirdek noktaların rasgele seçilmesi veya ilk kümelerdeki birimlerin rasgele ayrıştırılmasıdır.

Burada en çok kullanılan *K*-Ortalama Yöntemi üzerinde durulacaktır.

***K*-Ortalama Yöntemi**

K-ortalama yönteminde, her bir birim en yakın merkezli(ortalamalı) kümeye atanır. Bu yöntem aşağıdaki adımlardan oluşur:

Adım 1. Birimler *K* tane kümeye ayrılır ve oluşan kümelerin merkezleri(ortalamaları) hesaplanır.

Adım 2. Birimlerin listesinden, her bir birimi en yakın merkezli(ortalamalı) kümeye ata (Birimleri kümelere olan uzaklıkları için genelde Öklid uzaklığı kullanılır). Birimini kaybeden küme ve yeni birim alan kümeler için merkez değerleri(ortalamalar) yeniden hesaplanır.

Adım 3. Adım 2 yeniden atama olmayana kadar(yani kümelerdeki birimlerin yine bulunduğu kümede kalması) tekrarlanır.

Bununla birlikte birimleri rasgele *K* kümeye ayırmak yerine, *K* tane merkez belirlenerek de işe başlanılabilir. Bu durumda 1. Adıma gerek kalmaz.

Kümelere birimlerin son atanması; ilk kümelerin oluşturulması veya merkez noktalarının ilk seçimine bağlıdır. Bu durum işlem süresine etkilemektedir.

Örnek 32: *A*, *B*, *C* ve *D* ile gösterilen dört birimin, X_1 ve X_2 rasgele değişkenlerine ilişkin gözlem değerleri aşağıdaki gibidir. Bu 4 birimi *K*=2 kümeye ayırınız.

Birimler	Gözlem Değerleri	
	x_1	x_2
A	5	3
B	-1	1
C	1	-2
D	-3	-2

Çözüm 32: Amaç birimler *K*=2 kümeye ayrılırken, aynı kümedeki birimler, farklı kümedeki birimlere göre bir birine daha yakın olsun. *K*- ortalamalar yönteminde birimler başlangıçta rasgele gruplara ayrılabilir. *K*=2 olduğundan; bu dört birim, birinci küme *A* ve *B* birimlerinden ve ikinci küme *C* ve *D* birimlerden oluşacak biçimde rasgele olarak iki kümeye ayrıldığını kabul edelim. Adım 1'e göre oluşturulan bu kümelere ilişkin küme merkezleri

Kümeler	Küme Merkezleri (Ortalamaları)	
	$\overline{x}_1$	$\overline{x}_2$
(*AB*)	$\frac{5+(-1)}{2}=2$	$\frac{3+1}{2}=2$
(*CD*)	$\frac{1+(-3)}{2}=-1$	$\frac{-2+(-2)}{2}=-2$

biçiminde elde edilir. Adım 2 uygulanarak, her bir birimin küme merkezlerine olan uzaklıkları Oklid uzaklığına göre hesaplanır ve birimler en yakın olduğu kümeye atanır. Eğer bir birim ilk bulunduğu kümeden ayrılırsa, küme merkezleri yeniden hesaplamalıdır. Kare uzaklıkları;

$$d^2(A,(AB)) = (5-2)^2+(3-2)^2 = 10$$

$$d^2(A,(CD)) = (5-(-1))^2+(3-(-2))^2 = 61$$

A birimi (AB) kümesine daha yakındır.

$$d^2(B,(AB)) = ((-1)-2)^2+(1-2)^2 = 10$$

$$d^2(B,(CD)) = (-1-(-1))^2+(1-(-2))^2 = 9$$

B birimi (CD) kümesine daha yakındır.

$$d^2(C,(AB)) = (1-2)^2+((-2)-2)^2 = 17$$

$$d^2(C,(CD)) = (1-(-1))^2+((-2)-(-2))^2 = 4$$

C birimi (CD) kümesine daha yakındır.

$$d^2(D,(AB)) = ((-3)-2)^2+((-2)-2)^2 = 41$$

$$d^2(D,(CD)) = ((-3)-(-1))^2+((-2)-(-2))^2 = 4$$

D birimi (CD) kümesine daha yakındır.

Buradan yeni kümeler ve küme merkezleri (ortalamaları)

Kümeler	Küme Merkezleri (Ortalamaları)	
	$\overline{x}_1$	$\overline{x}_2$
(A)	5	3
(BCD)	$\frac{-1+1+(-3)}{3} = -1$	$\frac{1+(-2)+(-2)}{3} = -1$

olarak elde edilir. Birimlerin elde edilen bu kümelere göre uzaklıkları

$$d^2(A,(A)) = (5-5)^2+(3-3)^2 = 0$$

$$d^2(A,(BCD)) = (5-(-1))^2+(3-(-1))^2 = 52$$

A birimi (*A*) kümesine daha yakındır.

$$d^2(B,(A)) = ((-1)-5)^2+(1-3)^2 = 40$$

$$d^2(B,(BCD)) = (-1-(-1))^2+(1-(-1))^2 = 4$$

B birimi (*BCD*) kümesine daha yakındır.

$$d^2(C,(A)) = (1-5)^2+((-2)-3)^2 = 41$$

$$d^2(C,(BCD)) = (1-(-1))^2+((-2)-(-1))^2 = 5$$

C birimi (*BCD*) kümesine daha yakındır.

$$d^2(D,(A)) = ((-3)-5)^2+((-2)-3)^2 = 89$$

$$d^2(D,(BCD)) = ((-3)-(-1))^2+((-2)-(-1))^2 = 5$$

D birimi (*BCD*) kümesine daha yakındır.

Her birim kendi kümesinde kaldığı için, işlem burada sonlandırılır. Böylece *A*, *B*, *C* ve *D* ile gösterilen dört birim; (*A*) ve (*BCD*) olmak üzere iki kümeye ayrılmış olur.

Küme Sayısının Belirlenmesi

Uygun küme sayısına karar vermek için birkaç yol söz konusudur:

1. En pratik yollardan biri *k* küme sayısı ve *n* birim sayısı olmak üzere

   $$k \cong \sqrt{\frac{n}{2}}$$

   olarak belirlenir. Bu yöntem küçük örneklemlerde kullanılmaktadır.

2. Marriot tarafından önerilen yöntemde, *W* grup içi (Küme içi) kareler ve çapraz çarpımlar toplamı matrisi olmak üzere, küme sayısı

   $$M = k^2|W|$$

   eşitliği ile bulunan en küçük *M* değerini veren *k* sayısı küme sayısı olarak alınır. Burada

$$W = \sum_{l=1}^{k} \sum_{j=1}^{n_l} (\underline{x}_{lj} - \underline{\bar{x}}_l)(\underline{x}_{lj} - \underline{\bar{x}}_l)' \quad ,$$
$$= (n_1 - 1)S_1 + (n_2 - 1)S_2 + ... + (n_k - 1)S_k$$

$$S_l = \sum_{j=1}^{n_l} (\underline{x}_{lj} - \underline{\bar{x}}_l)(\underline{x}_{lj} - \underline{\bar{x}}_l)' \;\; ; \;\; l = 1, 2, ..., k \quad ,$$

$\underline{x}_{lj}$ l inci kümedeki j inci birime ait gözlem vektörü ve $\underline{\bar{x}}_l$ l inci kümedeki birimlerin gözlem değerlerinin ortalama vektörüdür.

3. Calinsky ve Harabazs tarafından önerilen yöntemde,

$$C = \frac{tr(B)/(k-1)}{tr(W)/(n-k)}$$

değerini en büyük yapan k değeri uygun küme sayısı olarak alınmaktadır. Burada B gruplar arası kareler ve çapraz çarpımlar toplamı matrisidir ve

$$B = \sum_{l=1}^{k} n_l (\underline{\bar{x}}_l - \underline{\bar{x}})(\underline{\bar{x}}_l - \underline{\bar{x}})'$$

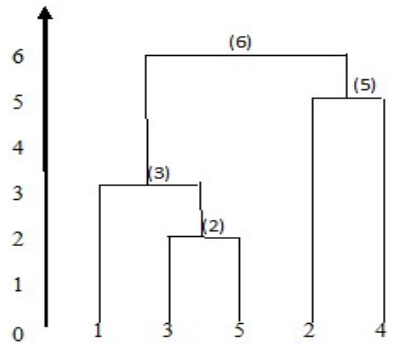
dır ve $\underline{\bar{x}}$ tüm kümelerdeki birimlerin gözlem değerlerinin genel ortalama vektörüdür.

4. Bunlardan başka uzaklık matrisi D' nin elemanlarının sıralanmasıyla elde edilen serilerden en büyük aralık değerlerine göre de küme sayısı pratik olarak belirlenebilir.

5. Ayrıca oluşturulan her bir küme, bir kitle olarak kabul edilip, bu kitlelerin ortalamaları arası farklılık olup olmadığı bakılarak da küme sayısı belirlenebilir. Burada Karşılaştırma yapmak için Hotelling T^2 istatistiği kullanılabilir.

Örnek 33: Daha önce verilen *n*=5 birimin Tek Bağlantı kümelenmesinden elde edilen uzaklıklara göre oluşan Dendogram'a göre küme sayısını belirleyiniz.

Çözüm 33:



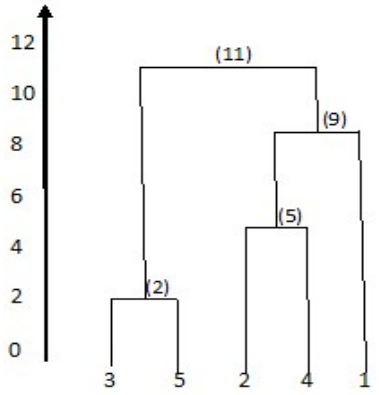
Dendogramdan uygun küme sayısı, uzaklıklar matrisinin değerlerinin sıralanmasıyla elde edilir.

Uzaklık	Kümeler	Küme sayısı
0	1,2,3,4,5	5
2	(35),1,2,4	4
3	(135),2,4	**3**
5	(135),(24)	2
6	(12345)	1

En büyük uzaklık artışı(2 birim); 3 kümeden, 2 kümeye düşerken gerçekleştiğinden, küme sayısı 3 dür. (2 birimlik artış 5 kümeden, 4 kümeye düşerken de gerçekleşmiştir. Toplamda 5 birim olduğu için, küme sayısının 5 alınması anlamlı olmayacaktır). Ancak küme sayısına bu yolla karar vermede problemin özelliği ve araştırıcının tecrübesi de önemlidir. Bu yöntem her zaman kolay uygulanamayabilir.

Örnek 34: Daha önce verilen *n*=5 birimin Tam Bağlantı kümelenmesinden elde edilen uzaklıklara göre oluşan Dendogram'a göre küme sayısını belirleyiniz.

Çözüm 34:



Uzaklık	Kümeler	Küme sayısı
0	1,2,3,4,5	5
2	(35),1,2,4	4
5	(35),(24),1	**3**
9	(35),(124)	2
11	(12345)	1

En büyük uzaklık artışı(4 birim); 3 kümeden, 2 kümeye düşerken gerçekleştiğinden, küme sayısı 3 dür.