SAKARYA ÜNİVERSİTESİ

Veri Madenciliği Uygulamaları

Hafta 10

Yrd. Doç.Dr. Nilüfer YURTAY

Kümeleme ve Hiyerarşik Yöntemler

10.1 Giriş

Kümeleme işlemi birbirine benzeyen nesnelerin aynı grupta toplanmasıdır. Bu modelde en büyük etken hangi kriterlere göre kümeleme yapılacağıdır. Bu işlem konu ile ilgili uzman tarafından tahmin edilir. Veriler kümeleme işleminde aynı sınıfta yer almalarına rağmen farklı gruplarda da yer alabilir. Nüfus bilimi ve astronomi alanında kullanımları yaygındır.

Kümeleme analizi veriyi anlamlı, yararlı yada hem anlamlı hem de yararlı gruplara(kümelere) ayırır. Eğer amaç anlamlı gruplar ise, bu durumda kümeler verinin doğal yapısını yakalamalıdır. Bazı durumlarda, kümeleme analizi başka amaçlar için yalnızca bir başlangıç noktasıdır. Kümeleme analizi bir çok alanda uzun bir süre için önemli bir rol oynadı. Örneğin, psikoloji ve diğer sosyal bilimler, biyoloji, istatistik, patern tanıma, bilgi çıkarma, makine öğrenmesi ve veri madenciliği[1].

Biyologlar uzun yıllar canlılar için bir hiyerarşik sınıflandırma çabası içinde oldular. Örnek olarak krallık, birlik, sınıf, düzen, aile, cins ve tür verilebilir. Biyologların kümeleme analizindeki ilk çalışmaların sınıflandırma yapılarını otomatik olarak elde edecek matematiksel taksonomi üzerinde yoğunlaşması şaşırtıcı değildir. Son zamanlarda, biyologlar sınıflandırma analizini büyük miktardaki gen bilgilerine uygulamaktadırlar. Aynı özellikteki genlerin bulunması gibi…

Örütbağ (World Wide Web) milyarlarca web sayfası içerir ve bir arama motoruna yapılacak bir sorgu binlerce sayfa geri döndürebilir. Kümeleme bu bilgilerin çeşitli gruplara ayrılmasında kullanılabilir böylece her grup sorgunun bir yönüne yönelik olur. Örnek olarak bir film sorgusu sonuçları eleştiri, fragman, yıldızlar ve tiyatrolar şeklinde kümelere ayırabilir. Her küme tekrardan kümelere ayrılabilir böylece kullanıcının sonuçları daha iyi irdelemesine yardımcı olabilir.

Yeryüzü iklimini anlamak, atmosfer ve okyanuslara yönelik çeşitli paternlerin bulunmasını gerektirir. Şu ana kadar, sınıflandırma analizi kutupsal bölgelerin atmosferik basınçlarına ilişkin paternlerin ve kara iklimine önemli etkisi bulunan okyanus alanlarının bulunmasında kullanılmıştır.

Bir hastalık veya sağlık durumu sık sık çeşitli varyasyonlar gösterir ve kümeleme analizi bu değişik çeşitlilikleri ortaya çıkarmada kullanılabilir. Örnek olarak kümeleme depresyonun değişik türlerinin belirlenmesinde kullanılmıştır. Kümeleme analizi aynı zamanda hastalıkların zaman ve mekanda dağılımı ile ilgili paternlerin ortaya çıkarılmasında da kullanılabilir.

Ticari işlemlerde Kümeleme, müşterileri daha küçük alt gruplara ayırmada, böylece fazladan analiz ve pazarlama aktiviteleri yürütmede kullanılabilir.

Küme prototipleri veri sıkıştırma için de kullanılabilirler. Özellikle, her bir küme için prototipleri içeren tablolar oluşturulur yani her bir prototipe onun tablodaki yerini gösteren bir indis değeri atanır.

Kümeleme, hedef değişken için sınıflama, kestirim, tahminleme gibi değerler bulmaya çalışmaz. Sınıflandırmadan farklı olarak daha önceden tanımlanmış sınıflara ve sınıf etiketlerine dayanmaz.

[1] *www.bilmuh.gyte.edu.tr/~htakci/vm/**kumeleme**_analizi.doc*

10.2 Hiyerarşik Kümeleme Yöntemleri

Kümeleme analizinin kullanılmasında benzerlik uzaklıklar dikkate alınarak yaralanılabilecek çok fazla alternatif ölçü ve yöntem bulunmaktadır. Örneğin sadece birimler arası uzaklıklar için Euclidyen, Kareli Euclidyen, Standardize Euclidyen, Manhattan Mahalanobis, Minkowski veya Canberra ölçüleri kullanılabilmektedir[2]. Bu da kümeleme analizinin uygulamada kullanılmasında dikkatli davranmayı zorunlu kılmaktadır.

Diğer çok değişkenli istatistik tekniklerinde önemli olan verilerin normalliği varsayımı, kümeleme analizinde çok önemli olmayıp, uzaklık değerlerinin normalliği yeterli görülmektedir.

Kümeleme analizinin aşamaları şu şekilde özetlenebilir:

- Birimler arasında var olan benzerliğin belirlenebilmesi için kullanılacak ölçülerin ve değişkenlerin belirlenmesi,
- Birimler arasındaki benzerliklerin belirlenmesinden sonra birimlerin kümelenmesi
- Oluşturulan kümelerin uygun olup olmadığının belirlenmesi
- Kümelerin uygun olarak elde edildiği varsayımı altında bunun istatistik geçerliliğinin ortaya konması, şeklinde sıralanabilir.

Kümeleme algoritmaları hiyerarşik (tekli bağlantı, tam bağlantı, ortalama bağlantı) ve hiyerarşik olmayan (k ortalamalı kümeleme) olarak ikiye ayrılır.

Tekli Bağlantı (Single Linkage): Bazen yakın komşu yaklaşımı olarakta adlandırılır. A kümesindeki bir kayıtla B kümesindeki bir kaydın minimum uzaklığını temel alır. Tekli bağlantı, uzun, yetersiz kümelerin oluşumuna ve kümelenmiş kayıtların heterojen olmasına sebep olabilir.

Tam Bağlantı (Complete Linkage): Bazen uzak komşu yaklaşımı olarakta adlandırılır. A kümesindeki bir kayıtla B kümesindeki bir kaydın maximum uzaklığını temel alır. Tam bağlantı ile daha sık, küresel kümelerin oluşması sağlanır.

Ortalama Bağlantı (Average Linkage): Benzeyen ya da benzemeyen kayıtlar için uç değerlerdeki küme bağlantı kriterinin bağlılığını azaltmak için düzenlenmiştir. A kümesi ve B kümesi arasındaki tüm kayıtların ortalama uzaklığıdır. Sonuçta oluşan kümeler, küme içi değişimin yaklaşık olarak eşit olmasına yol açar.

k Ortalamalı Kümeleme (k Means Clustering): Küme içinde objelerin ortalamaları üzerinde kurulur. Hata kareler toplamını (SSE) minimum yapacak k tane küme alınır. Küme sayısı keyfi belirlenir. Veriyi ortalaması en yakın olan kümeye atar. Sonra yeniden ortalama hesaplanır. Bu şekilde ortalama sürekli güncellenir, ta ki fark oluşmayıncaya kadar.

İlişkilendirme (Association): İlişkilendirme algoritması, bir ilişkide bir niteliğin aldığı değerler arasındaki bağımlılıkları, diğer niteliklere göre gruplama yapılmış verileri kullanarak bulur. Keşfedilen ilişkiler (örüntüler) örneklemde sıklıkla birlikte geçen nitelik değerleri arasındaki ilişkiyi gösterir.

[2] http://iletisim.atauni.edu.tr/eisemp/html/tammetinler/172%20.pdf

10.3 Tek Bağlantılı Hiyerarşik Kümeleme Yöntemi

Bazen yakın komşu yaklaşımı olarakta adlandırılır. A kümesindeki bir kayıtla B kümesindeki bir kaydın minimum uzaklığını temel alır. Tekli bağlantı, uzun, yetersiz kümelerin oluşumuna ve kümelenmiş kayıtların heterojen olmasına sebep olabilir.

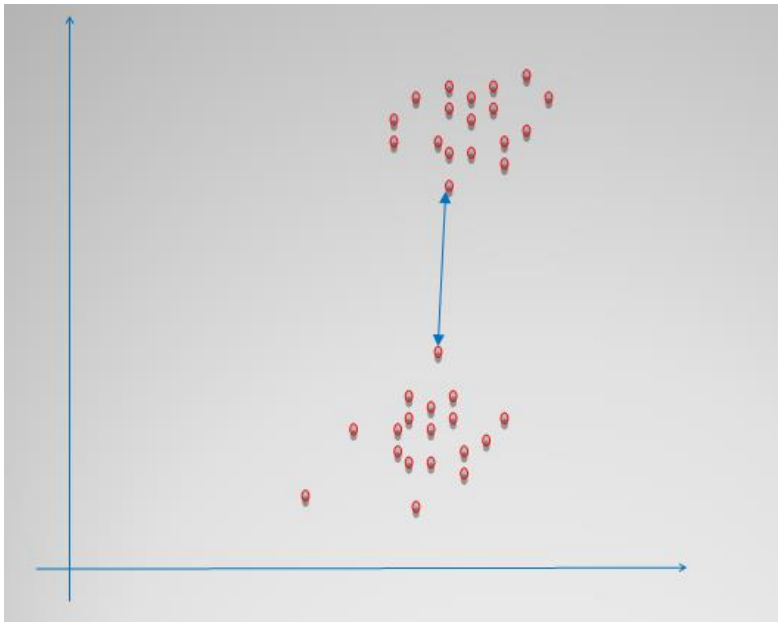
Bu yöntemde başlangıç aşamasında tüm gözlem değerleri birer küme olarak ele alınır ve adım adım bu kümeler birleştirilerek yeni kümeler elde edilir.

Gözlem değerleri arasındaki minimum uzaklıkların hesaplanabilmesi için öklid uzaklık formülünden yaralanılabilir.

$$uzaklık\,(i,\,j) = \sqrt{\sum_{k=1}^{p}(x_{ij} - x_{jk})^2}$$

En düşük uzaklık seçilerek bu uzaklıkla ilgili elemanlar birleştirilip yeni bir küme elde edilir. Daha sonra uzaklıklar yeniden hesaplanır.

Birden fazla gözlem değerine sahip olan iki küme arasındaki uzaklığın belirlenmesi gerektiğinde farklı yollar izlenebilecektir. Bu gözlemler arasında birbirine en yakın olanların uzaklığı iki kümenin birbirine olan uzaklığı olarak değerlendirilir(şekil 10.1).

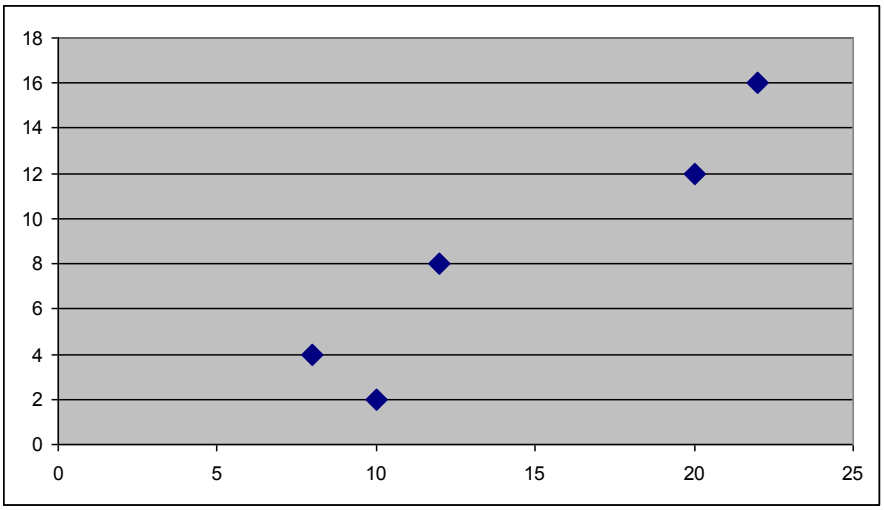


Şekil 10.1 Tek gözlemli olmayan İki küme arasındaki uzaklığın hesaplanması

10.4 Örnek Uygulama

Tablo değerlerinden hareketle Tek Bağlantılı Hiyerarşik Kümeleme Yöntemi (en yakın komşu) algoritmasını kullanarak kümeleme yöntemini uygulamak istersek.

hasta no	ilk ay için migren atak sayısı	atak süresi
1	8	4
2	12	8
3	10	2
4	20	12
5	22	16



İlk olarak uzaklık tablosunu oluşturalım:

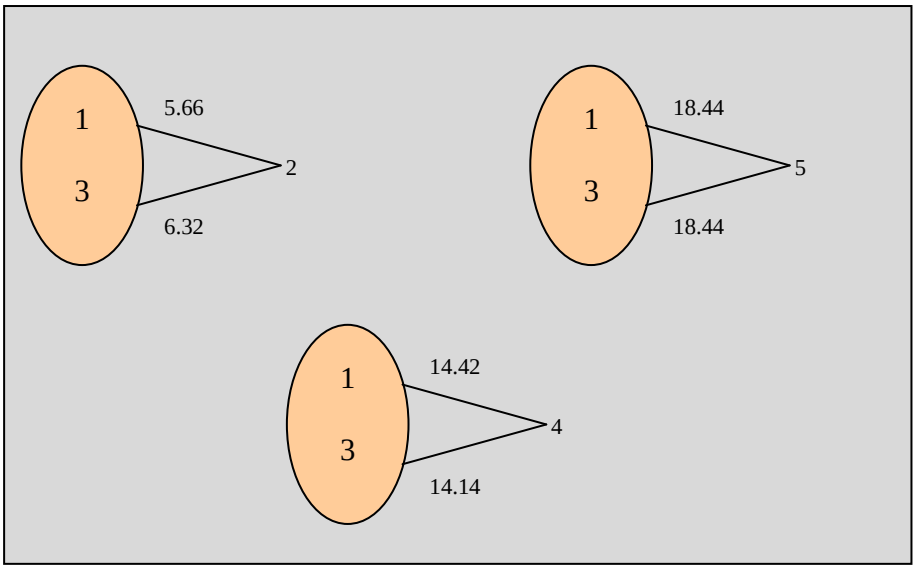
$$uzak(i,j) = \sqrt{\sum_{k=1}^{p}(x_{ij} - x_{jk})^2}$$

$$uzak(1,2) = \sqrt{(8-12)^2 + (4-8)^2} = 5.66$$
$$uzak(1,3) = \sqrt{(8-10)^2 + (4-2)^2} = 2.83$$
$$uzak(1,4) = \sqrt{(8-20)^2 + (4-12)^2} = 14.42$$
$$uzak(1,5) = \sqrt{(8-22)^2 + (4-16)^2} = 18.44$$
$$uzak(2,3) = \sqrt{(12-10)^2 + (8-2)^2} = 6.32$$
$$uzak(2,4) = \sqrt{(12-20)^2 + (8-12)^2} = 8.94$$
$$uzak(2,5) = \sqrt{(12-22)^2 + (8-16)^2} = 12.81$$
$$uzak(3,4) = \sqrt{(10-20)^2 + (2-12)^2} = 14.14$$
$$uzak(3,5) = \sqrt{(10-22)^2 + (2-16)^2} = 18.44$$
$$uzak(4,5) = \sqrt{(20-22)^2 + (12-16)^2} = 4.47$$

Elde edilen uzaklık matrisi aşağıdaki gibidir:

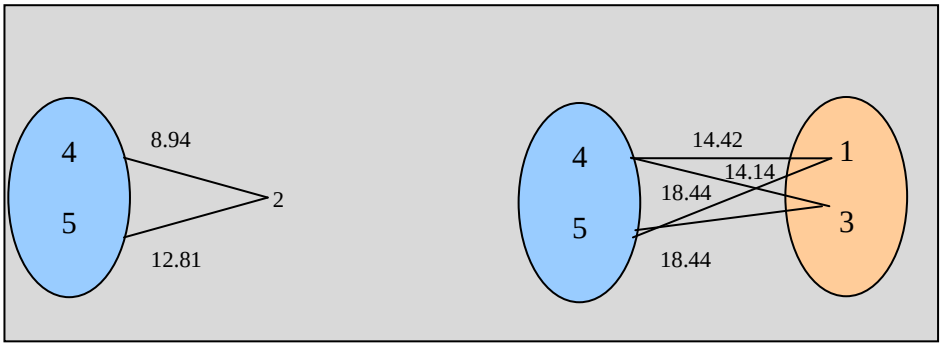
Hasta no	1	2	3	4	5
1					
2	5.66				
3	**2.83**	6.32			
4	14.42	8.94	14.14		
5	18.44	12.81	18.44	4.47	

En düşük uzaklık 2.83 olup bu değere sahip 1 ve 3 nolu gözlemler birleştirilerek {1,3} kümesini elde ederiz. Yapılacak olan {1,3} kümesi ile diğer gözlemler arasındaki uzaklıkları bulmaktır:



Hasta no	{1,3}	2	4	5
{1,3}				
2	5.66			
4	14.14	8.94		
5	18.44	12.81	**4.47**	

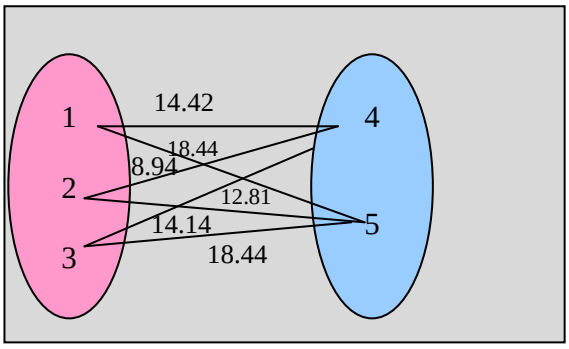
En düşük uzaklık 4.47 olup bu değere sahip 5 ve 4 nolu gözlemler birleştirilerek {4,5} kümesini elde ederiz. Yapılacak olan {4,5} kümesi ile diğer gözlemler arasındaki uzaklıkları bulmaktır:



Yeni uzaklık tablosu yazılabilir:

Hasta no	{1,3}	2	{4,5}
{1,3}			
2	**5.66**		
{4,5}	14.14	8.94	

En düşük uzaklık 5.66 olup bu değere sahip 2 ve {1,3} nolu gözlemler birleştirilerek {1,2,3} kümesini elde ederiz. Yapılacak olan {1,2,3} kümesi ile diğer gözlemler arasındaki uzaklıkları bulmaktır:



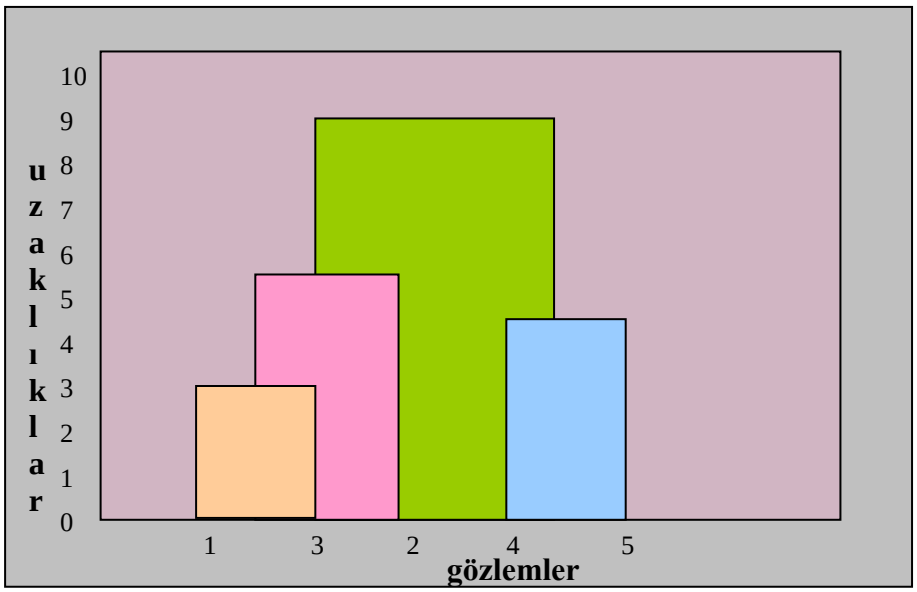
Yeni uzaklık tablosu yazılabilir:

Hasta no	{1,2,3}	{4,5}
{1,2,3}		
{4,5}	**8.94**	

Şimdi yapılması gereken çalışma, sonuç kümenin tanımlamasıdır:

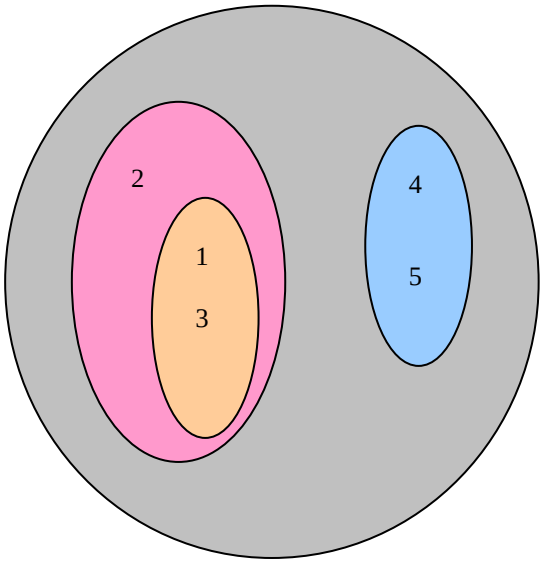
Uzaklıklar	Kümeler
2.83	{1,3}
4.47	{4,5}
5.66	{1,2,3}
8.94	{1,2,3,4,5}

Şimdi kümeleme çalışmamıza ait dendrogramı çizebiliriz (şekil 10.2):



Şekil 10.2 Dendrogram sonucu

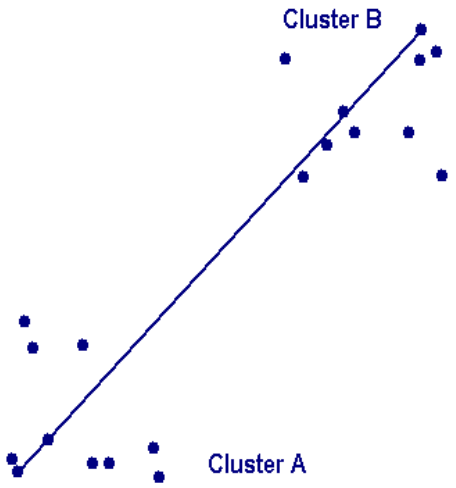
Çalışmamızla ilgili son olarak küme grafiğini de çizelim:



Şekil 10.3 Oluşan Kümeler

10.5 Tam Bağlantılı Hiyerarşik Kümeleme Yöntemi

Bazen uzak komşu yaklaşımı olarakta adlandırılır. A kümesindeki bir kayıtla B kümesindeki bir kaydın maximum uzaklığını temel alır. Tam bağlantı ile daha sık, küresel kümelerin oluşması sağlanır (şekil 10.4).



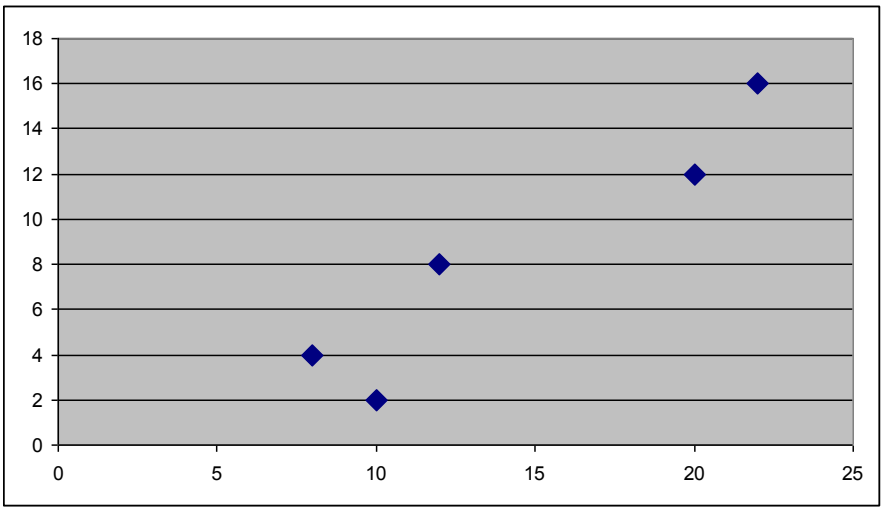
Şekil 10.4 Tam Bağlantılı Hiyerarşik kümeleme için yaklaşım

Tek bağlantılı hiyararşik kümeleme yönteminden farklı olarak uzaklıklar arasından en büyük olan seçilerek işlemler gerçekleştirilir.

10.6 Örnek Uygulama

Tablo değerlerinden hareketle Tam Bağlantılı Hiyerarşik Kümeleme Yöntemi (en uzak komşu) algoritmasını kullanarak kümeleme yöntemini uygulamak istersek.

hasta no	ilk ay için migren atak sayısı	atak süresi
1	8	4
2	12	8
3	10	2
4	20	12
5	22	16

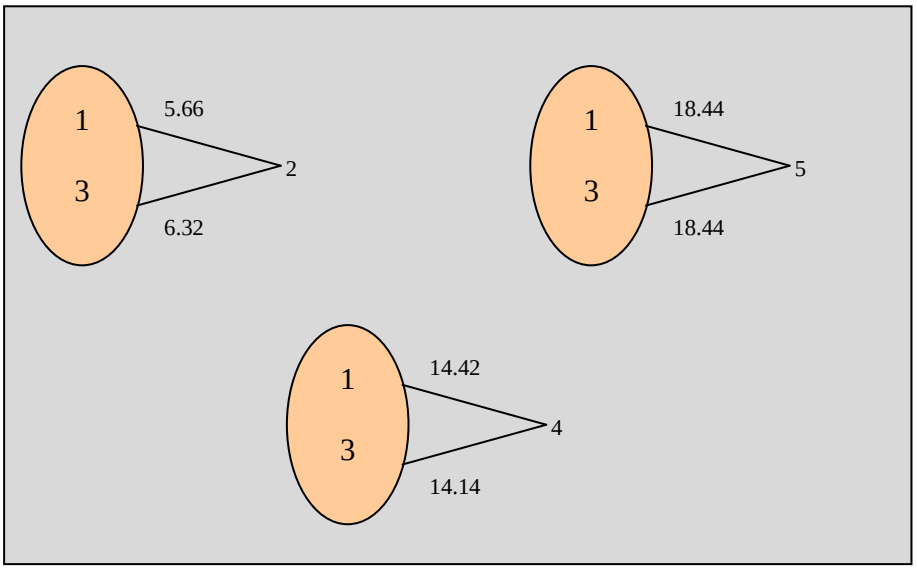


Yine ilk olarak öklid bağıtısı ile uzaklıklar hesaplanırsa aşağıdaki tablo bulunacaktır:

Hasta no	1	2	3	4	5
1					
2	5.66				
3	**2.83**	6.32			
4	14.42	8.94	14.14		
5	18.44	12.81	18.44	4.47	

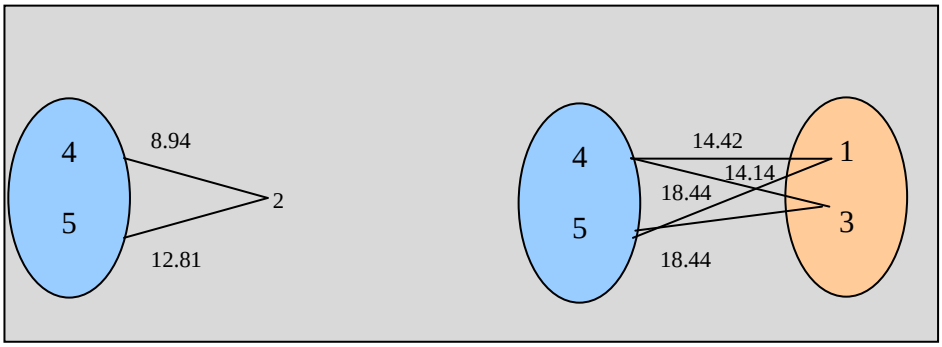
İlk aşamada tek Bağlantılı Hiyerarşik Kümeleme Yönteminde olduğu gibi en düşük hücre belirlenir:

En düşük uzaklık 2.83 olup bu değere sahip 1 ve 3 nolu gözlemler birleştirilerek {1,3} kümesini elde ederiz. Yapılacak olan {1,3} kümesi ile diğer gözlemler arasındaki uzaklıkları bulmaktır. Bunu yaparken de birbirine en uzak olan gözlemler seçilecektir:



Hasta no	{1,3}	2	4	5
{1,3}				
2	6.32			
4	14.42	8.94		
5	18.44	12.81	**4.47**	

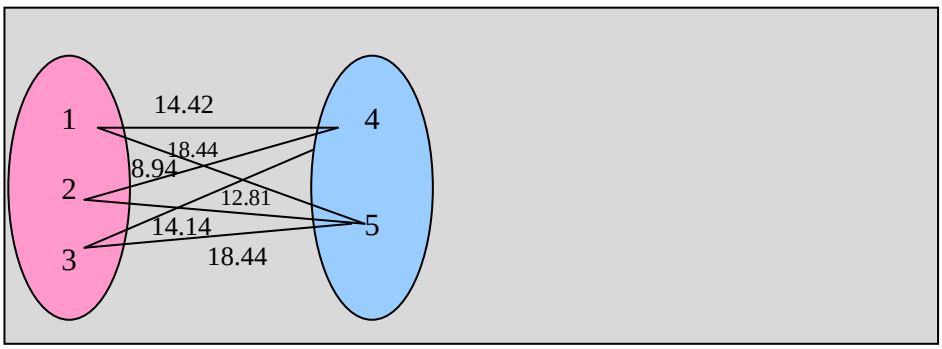
En düşük uzaklık 4.47 olup bu değere sahip 5 ve 4 nolu gözlemler birleştirilerek {4,5} kümesini elde ederiz. Yapılacak olan {4,5} kümesi ile diğer gözlemler arasındaki uzaklıkları bulmaktır:



Yeni uzaklık tablosu yazılabilir:

Hasta no	{1,3}	2	{4,5}
{1,3}			
2	**5.66**		
{4,5}	18.44	12.81	

En düşük uzaklık 5.66 olup bu değere sahip 2 ve {1,3} nolu gözlemler birleştirilerek {1,2,3} kümesini elde ederiz. Yapılacak olan {1,2,3} kümesi ile diğer gözlemler arasındaki uzaklıkları bulmaktır:



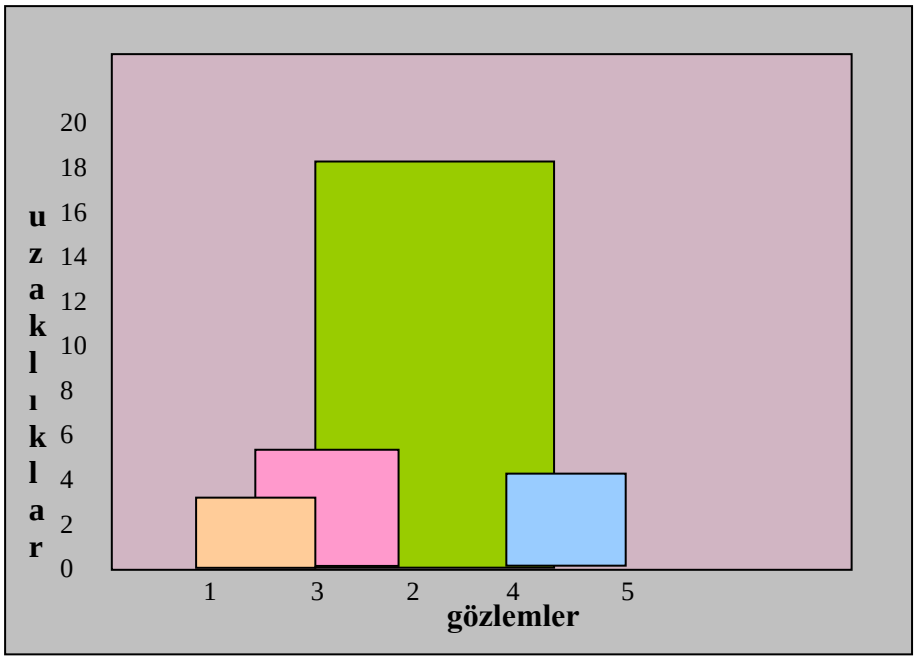
Yeni uzaklık tablosu yazılabilir:

Hasta no	{1,2,3}	{4,5}
{1,2,3}		
{4,5}	**18.44**	

Şimdi yapılması gereken çalışma, sonuç kümenin tanımlamasıdır:

Uzaklıklar	Kümeler
2.83	{1,3}
4.47	{4,5}
5.66	{1,2,3}
18.44	{1,2,3,4,5}

Şimdi kümeleme çalışmamıza ait dendrogramı çizebiliriz (şekil 10.5):



Şekil 10.5 Dendrogram