www.tubiad.org
ISSN:2148-3736
El-Cezerî Fen ve Mühendislik Dergisi
Cilt: 3, No: 2, 2016 (315-323)

El-Cezerî Journal of Science and Engineering
Vol: 3, No: 2, 2016 (315-323)

ECJSE

Makale / Research Paper

K-Means Algoritması İle Otomatik Kümeleme

Buket ÇOLAK[1], Zehra DURDAĞ[2], Pakize ERDOĞMUŞ[3]

[1,2,3]Mühendislik Fakültesi, Bilgisayar Mühendisliği ABD, Düzce Üniversitesi, Düzce - Türkiye

Geliş/Received: 07.11.2015 **Düzeltme/Revised:** 26.11.2015 **Kabul/Accepted:** 26.11.2015

Özet: K-Means kümeleme algoritması verileri K giriş parametre sayısı kadar kümeye bölmektedir. Bu çalışmanın amacı ise kümelemeyi otomatik hale getirmek ve dışarıdan K parametresinin girilmesine gerek kalınmadan verileri uygun küme sayısınca kümelere yerleştirmektir. Geliştirilen otomatik K-Means algoritması sayısal veriler ve görüntüler üzerinde test edilmiş ve başarılı sonuçlara ulaşılmıştır.

Anahtar kelimeler: K-means Kümeleme, Benzer Nesneler, Otomatik Kümeleme.

Automatic Clustering With K-Means

Abstract: K-means clustering algorithm groups the input data to K clusters. The object of this study is to automatize the clustering and grouping the data to optimum clusters without taking K parameter. Developed automatic K-means algorithm has been tested on data and some images and reached successful results.

Keywords: K-Means Clustering, Similar Objects, Auotomatic Clustering

1. Giriş

Kümeleme analizi, veri tabanlarındaki verilerin gruplar veya kümeler altında toplanarak, benzer özelliklere sahip nesnelerin bir araya gelmesini sağlayan bir veri madenciliği tekniğidir. Veri sayısı ve özellik sayısı arttıkça verileri kümelemek de zorlaşmaktadır. Kümeleme analizi tıpta, mühendislikte, ekonomide ve ziraatta olmak üzere çok çeşitli alanlarda kullanılmıştır. Bu konuda yapılmış çok sayıda çalışma Koltan ve arkadaşları tarafından sunulmuştur [1]. Giray ve arkadaşları çalışmalarında Avrupa Ülkelerini intihar oranlarına göre Fuzzy c-means ile sınıflandırmışlar [2]. Aşan, kredi kartı kullanan banka müşterilerinin sosyo-ekonomik özellikleri bakımından gruplandırmasını kümeleme analizi ile gerçekleştirmiştir. Müşterileri cinsiyet, yaş, kredi kartı türü gibi değişkenlere göre de gruplayarak, ne tür müşteri grubuna gideceklerini tespit etmeye çalışmışlardır[3]. Çelik, Türkiye'deki illeri sağlık göstergelerine göre sınıflandırmıştır[4]. Şişeci ve arkadaşları ise k-means algoritması ile resimleri alt bloklara ayırarak segmentasyon yapmışlar ve diğer yöntemlere göre daha başarılı sonuçlar almışlardır[5].

Bu makaleye atıf yapmak için

Çolak, B., Durdağ, Z., Edoğmuş, P., "K-Means Algoritması İle Otomatik Kümeleme" El-Cezerî Fen ve Mühendislik Dergisi 2016, 3(2);315-323.

How to cite this article

Çolak, B., Durdağ, Z., Edoğmuş, P., "Automatic Clustering With K-Means" El-Cezerî Journal of Science and Engineering, 2016, 3(2);315-323.

Bu alanda kullanılmak üzere çok çeşitli kümeleme algoritması geliştirilmiştir. En temel gözetimsiz kümeleme algoritmaları k-means ve fuzzy c-means algoritmasıdır[6].

Bu çalışmada küme sayısı belli değil iken verilerin k-means ile nasıl sınıflandırılacağı üzerinde durulmuş ve sınıf sayısı belirli olmayan verilerde optimum sınıf sayısının belirlenerek k-means ile sınıflandırılması amaçlanmıştır.

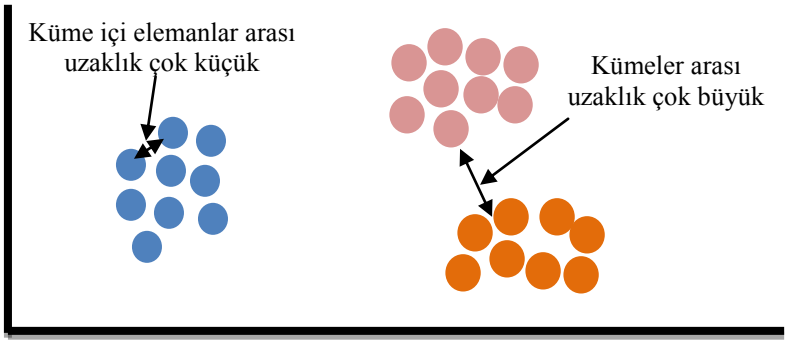
2. Yöntem

2.1 K-Means Kümeleme Yöntemi

Kümeleme, en basit tanımıyla benzer özellik gösteren verilerin kendi aralarında gruplara ayrılmasıdır. Kümeleme analizinde genel amaç küme içi homojenliği, kümeler arası heterojenliği sağlamaktır. Bu da benzer bireylerin aynı kümede toplanmasıyla sağlanabilir. Kümeleme problemi de bir optimizasyon problemi olup, küme elemanlarının küme ortalamasından uzaklıklarının toplamının minimizasyonu ile optimum kümeleme gerçekleştirilmiş olur. Bireylerin benzerlikleri uzaydaki konumları ile ilgilidir. Uzaydaki konumları itibari ile birbirine uzaklıkları daha az olan bireyler aynı kümede toplanmış olacaklardır.

Literatürde pek çok kümeleme algoritması bulunmaktadır. Kullanılacak olan kümeleme algoritmasının seçimi, amaca ve veri tipine bağlıdır. Genel olarak başlıca kümeleme yöntemleri; Bölme yöntemleri (Partitioning methods), Hiyerarşik yöntemler (Hierarchical methods), Yoğunluk tabanlı yöntemler (Density-based methods), Izgara tabanlı yöntemler (Grid-based methods), Model tabanlı yöntemler (Model-based methods) olarak sınıflandırılabilir [7].

Birimler arasındaki uzaklıkları hesaplamak için en sık kullanılan uzaklık ölçüleri Minkowski, Öklid (Euclide), Pearson, Manhattan (City-Blok), Mahalanobis, Hotelling T^2 ve Canberra Uzaklığı'dır [8]. Bu çalışmada uzaklık hesabı için Öklid uzaklığı kullanılmıştır.



Şekil 1. Küme Yapısı (Cluster Structure)[9]

Kümeleme işleminde küme içindeki nesneler arası uzaklık çok küçükken, kümeler arası uzaklık çok büyüktür [9]. En yaygın kullanılan gözetimsiz öğrenme yöntemlerinden biri olan K-means algoritması her verinin sadece bir kümeye ait olabilmesine izin veren keskin bir kümeleme algoritmasıdır [10].

K-means algoritmasının genel mantığı n adet veri nesnesinden oluşan bir veri setini, k adet giriş parametresi sayısı kadar kümeye bölümlemektir. Amaç, gerçekleştirilen bölümleme işlemi sonunda elde edilen kümelerin, küme içi benzerliklerinin maksimum ve kümeler arası benzerliklerinin minimum olmasını sağlamaktır. Bölümleyici kümelemeli yöntemlerden olan

K-Means algoritması sürekli olarak kümelerin yenilendiği ve en uygun çözüme ulaşana kadar devam eden döngüsel bir algoritmadır [6,11].

K-means yönteminin performansını k küme sayısı, başlangıç olarak seçilen küme merkezlerinin değerleri ve benzerlik ölçümü kriterleri etkilemektedir [12].

Küme sayısının belirlenmesi konusunda son yıllarda yoğun çalışmalar yapılmaktadır. Küme sayısının belirlenmesinde kullanılan en pratik yol (1) nolu eşitlik ile ifade edilir [13]. Ancak veri sayısının çok büyük olması durumunda pratik değildir.

$$k = \sqrt{\frac{n}{2}} \qquad (1)$$

n: kümelenecek birey sayısı

K-means kümeleme yönteminin değerlendirilmesinde en yaygın olarak karesel hata kriteri SSE kullanılır. En iyi sonucu en düşük SSE değerine sahip kümeleme verir. Nesnelerin bulundukları kümenin merkez noktalarına olan uzaklıklarının karelerinin toplamı (1) nolu eşitlik ile hesaplanmaktadır.

$$SSE = \sum_{i=1}^{K} \sum_{x \in C_i} dist^2(m_i, x) \qquad (2)$$

x: C_i kümesinde bulunan bir nesne, mi: C_i kümesinin merkez noktası

Algoritma, karesel-hata fonksiyonunu minimize edecek şekilde k kümeyi belirlemeye gayret eder. K-means algoritması, algoritmaya kullanıcı tarafından verilen k parametresi ile n tane veriden oluşan veri setini k adet kümeye böler [14].

K-means algoritmasının işlem basamakları şöyledir:

1.Adım: İlk olarak küme merkezleri belirlenir. Bunun için iki farklı yol vardır. Birincisinde nesneler arasından küme sayısı olan k adet rasgele nokta seçilir veya merkez noktalar tüm nesnelerin ortalaması alınarak belirlenir.

2.Adım: Her nesnenin seçilen merkez noktalara olan uzaklığı hesaplanarak tüm nesneler k adet kümeden kendilerine en yakın olan kümeye yerleştirilir.

3.Adım: Oluşan kümelerin yeni merkez noktaları o kümedeki tüm nesnelerin ortalama değeri ile değiştirilir.

4.Adım: Merkez noktalar değişmeyene kadar 2. ve 3. adımlar tekrarlanır.

2.2 Geliştirilen Yöntem

Kümeleme yöntemlerinden en yaygın olarak kullanılan k means algoritmasında kümeleme işleminin gerçekleştirilebilmesi için küme sayısının önceden belirlenmiş olması gerekmektedir. K means algoritmasında küme sayısının belirlenmesi işlemi ise rastgele olarak veya uzman bilgi ile belirlenir. Belirlenen küme sayısı k means algoritmasının doğruluğunu ve performansını doğrudan etkilemekte olduğu için çalışmamızda küme sayısının otomatikleştirilmesi ele alınmıştır.

N adet veri için düşünülecek olursa küme sayısı 1-N arası bir değer olacaktır. En iyi ihtimalle kümeleme işlemlerinde verilerin hepsi bir gruba ait olabilir veya en kötü ihtimalle verilerin tamamı farklı gruplarda yer alabilir. Bu durumda kmeans algoritmasının otomasyonu 1-N arası olan k değerini belirlemekten ibarettir. Çalışmada bu problem bir optimizasyon problemi olarak ele alınmıştır. Denklem 2'deki verilerin küme merkezlerinden Öklid uzaklıklarının toplamı amaç fonksiyonu olarak ele alınmıştır. "Kmeans karesel hata toplamını minimize eden k adet kümeye ayırıyor ise, bu karesel toplam uygun küme sayısında en minimum değerini alacaktır."

Otomatik kmeans algoritmasının akış şeması aşağıda verilmiştir.
Test verisi olarak Matlab'ın iris veri seti kullanılmıştır. 150 adet 4 farklı özellik içeren bu veri seti üç farklı küme içermektedir. Verilerin hangi kümeye ait oldukları da 5. sütun olarak veri setinde verilmiştir. Bu veri seti setosa, versicolor ve virginica olmak üzere iris çiçeğinin üç farklı türüne aittir. Her bir türden 50 tanesine ait sepal uzunluk, sepal genişlik, petal uzunluk, ve petal genişlik değerleri verilmiştir.
Algoritmanın akış şeması aşağıda Şekil 2'de verilmiştir.

1.Başla
2. k=1, Fmine=0;tol belirle;
3. k-means çalıştır. Fmin=ToplamHataKare;
4. Ebagil=abs(Fmin-Fmine)/Fmin>tol ise k=k+1; Fmine=Fmin; 3. Adıma git.
5. Dur

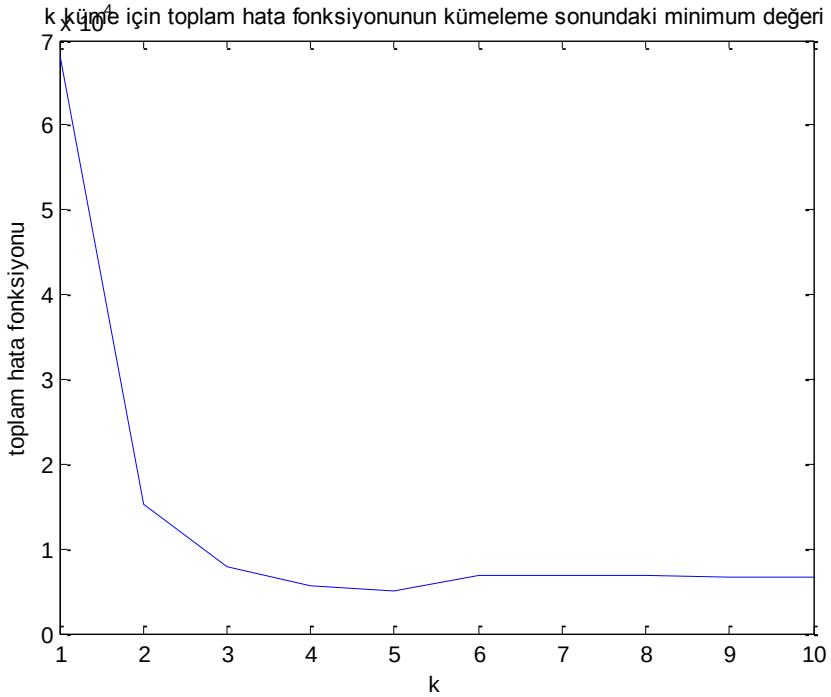
Şekil 2. Otomatik kmeans algoritması akış şeması

3. Bulgular

3.1. K-means Algoritmasının Sayısal Veriler için Sonuçları

Şekil 3.'te iris veri setinin k=1'den k=10 'e kadar çalıştırılması sonucu elde edilen toplam hata kare fonksiyonlarının kümeleme sonucundaki en minimum değerleri görülmektedir.

Şekil 3'te görüleceği üzere gerçek küme sayısından büyük küme değerlerinde artık hata fonksiyonu çok yavaş azalmaktadır.



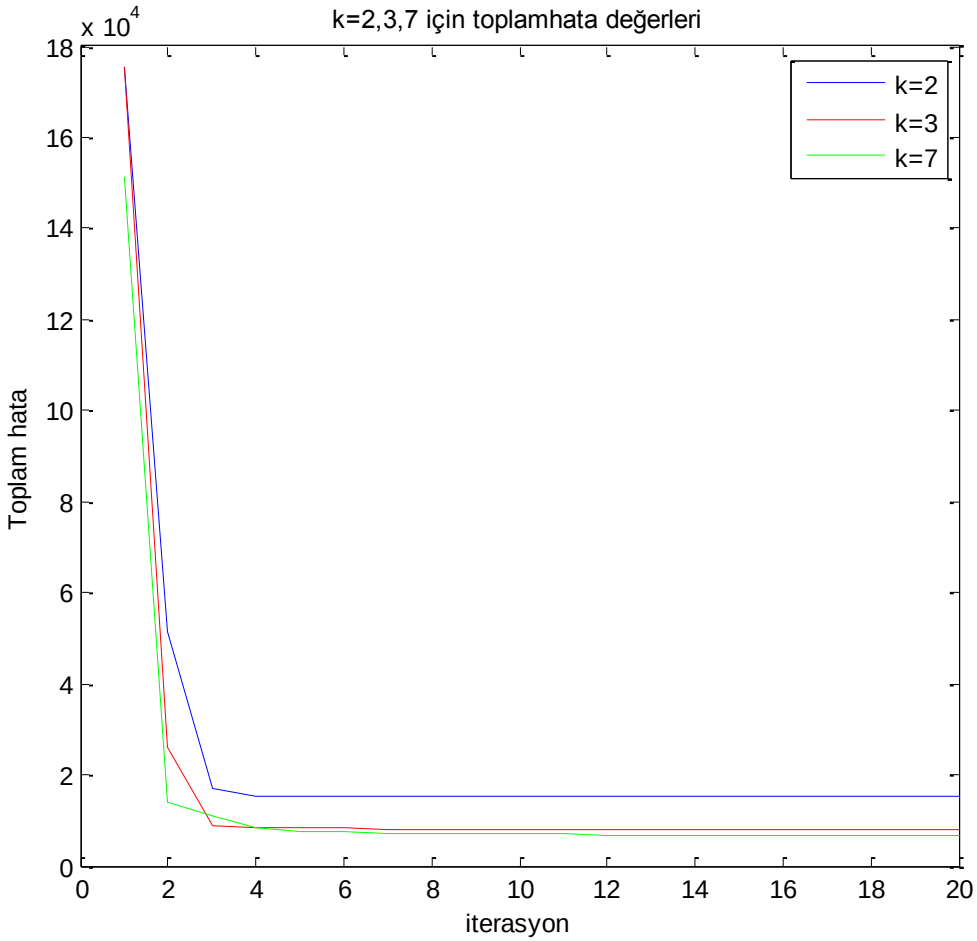
Şekil 3. Küme sayısına göre toplam hata fonksiyonunun değişimi

Tablo 1'de ise k(1-10) arası için bulunan toplam hata fonksiyonları ve bağıl hatalar verilmiştir.

Tablo 1. K küme sayısına göre minimum toplam hata değerleri

k	**ToplamHata**	**Bağıl Hata**	**Log(ToplamHata)**	**log Toplamın Bağıl Hatası**
1	68137,06	-	4,833383	
2	15234,8	3,472463	4,182837	0,155528
3	**7885,567**	**0,931985**	**3,896833**	**0,073394**
4	5725,601	0,377247	3,757821	0,036993
5	4984,981	0,14857	3,697664	0,016269
6	6872,671	0,274666	3,837126	0,036345
7	6833,895	0,005674	3,834668	0,000641
8	6760,238	0,010896	3,829962	0,001229
9	6733,542	0,003965	3,828244	0,000449
10	67334,708	0,000173	3,828319	1,97E-05

Şekil 4'de ise 20 iterasyon için k=2, k=3, k=7 için toplam hata fonksiyonu değerleri verilmiştir.



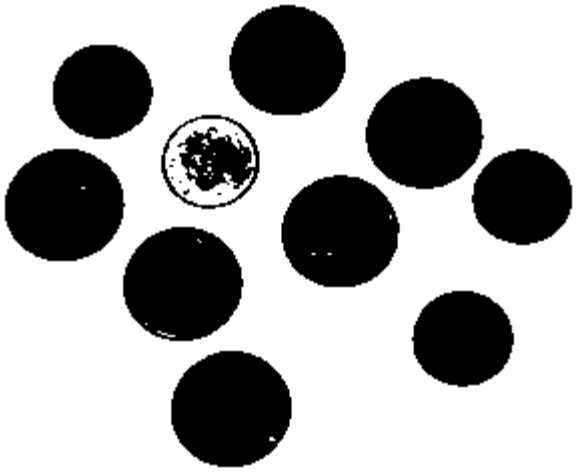
Şekil 4. K küme sayılarına göre toplam hata değişimi

3.1. K-means Algoritmasının Görüntüsel Veriler için Sonuçları

Algoritma Matlab'da yer alan test resimleri üzerinde de uygulanmış ve bulunan otomatik küme sayısı ve kümelenmiş resimler Şekil 5 ile Şekil 12 arasında sunulmuştur.

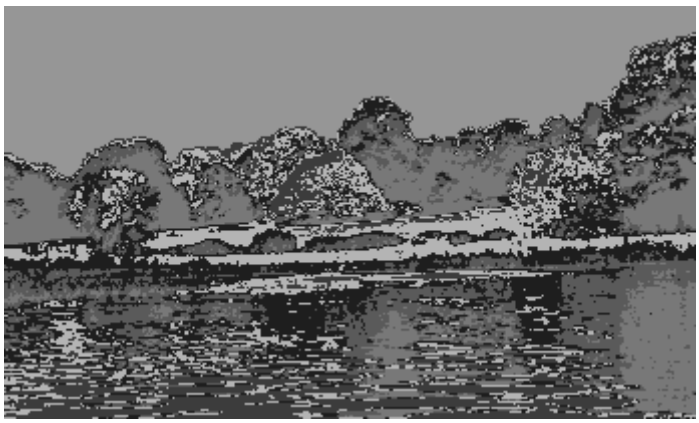
Coins.png resmi

Şekil 5. Bozuk paralar test resmi Matlab©

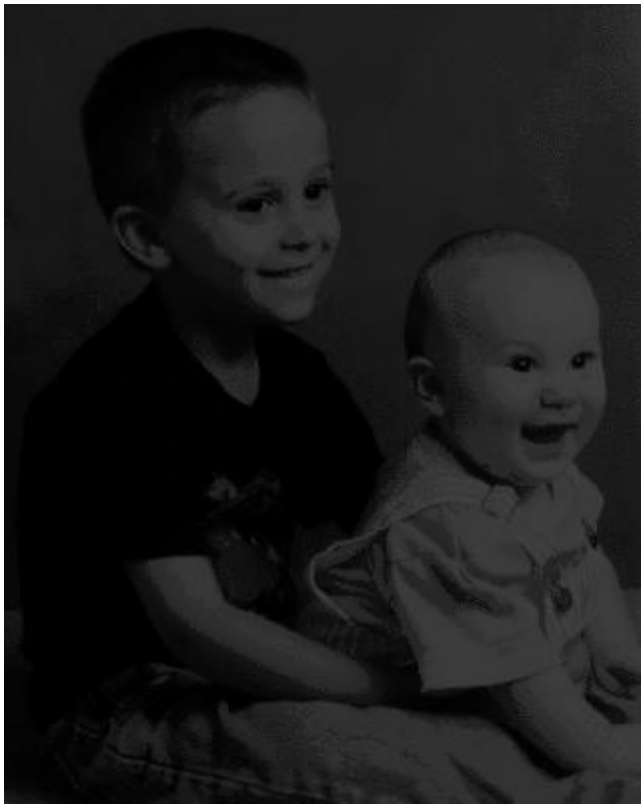


Şekil 6. Otomatik küme sayısınca renklendirilmiş bozuk paralar(k=2)

Şekil 7. Autumn.tif test resmi Matlab©

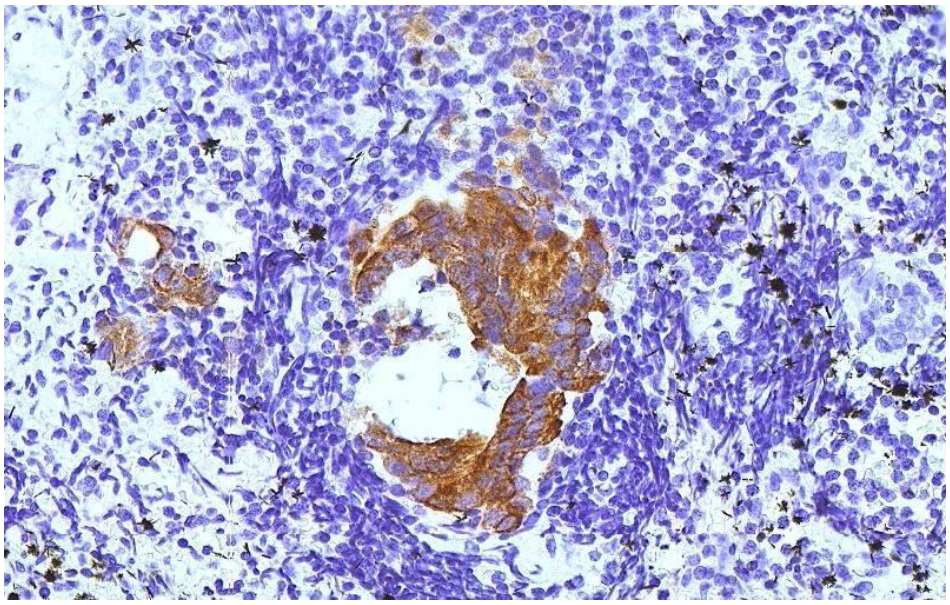


Şekil 8. Kümelenmiş Autumn.tif test resmi(k=6)

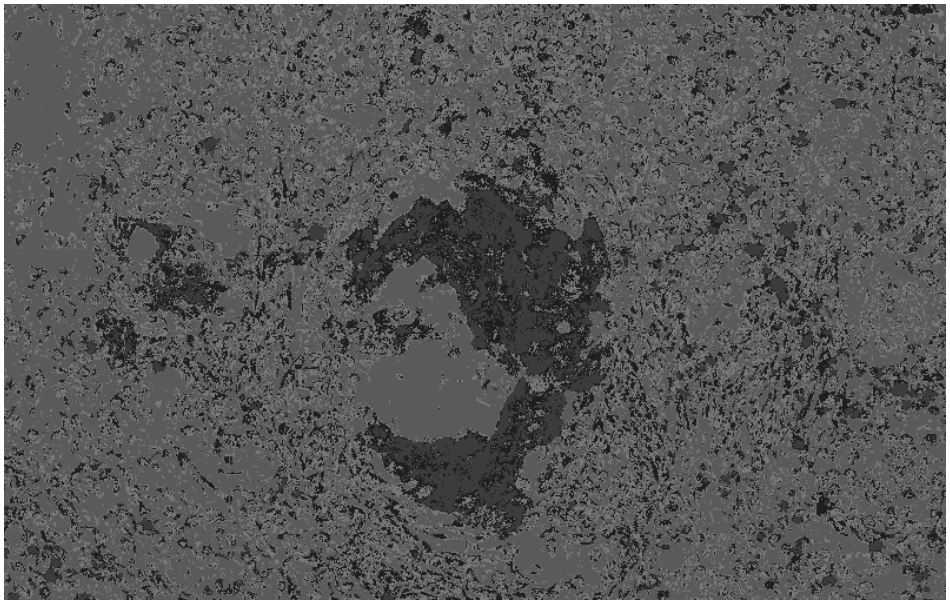


Şekil 9. Kids.tif test resmi Matlab©

Şekil 10. Kümelenmiş resim(k=4)



Şekil 11 . Tissue.tif test resmi Matlab©



Şekil 12. Kümelenmiş test resmi(k=4)

4. Sonuç

Kümeleme analizi, çok boyutlu uzayda birbirine yakın olan gözlemlerden meydana gelen grupları veya kümeleri bulmayı amaçlamaktadır. Diğer bir ifade ile analiz, örneklem verilerini gözlemlerin benzerliklerine göre en uygun kümelere ayırmaktadır. Kümeleme Analizi, kümelerin sayısına veya küme yapılarına ilişkin herhangi bir varsayımda bulunmaz. Bu çalışmada küme sayısına ilişkin bir varsayım yaparak kümeleme yapan bir algoritma geliştirilmiştir. Otomatik k-means algoritması en uygun k değerini tesbit ederek kümeleme yapmaktadır. Uygun küme değerinin belirlenmesi doğru sınıflandırma için oldukça önemlidir. Bankacılık sektöründe kaç farklı müşteri profilinin olduğu, resimde kaç farklı bölge olduğu, tıpta kaç farklı hasta tipi olduğu doğru küme sayısı ile tesbit edilebilir. Bu çalışma diğer kümeleme algoritmalarından fuzzy c-means algoritmasına da adapte edilebilir. Daha karmaşık veri setleri üzerinde de denenerek kümeleme performansları tesbit edilir.

Kaynaklar

[1] Koltan Ş, Patır S.(2011). Kümeleme Analizi ve Pazarlamada Kullanımı, Akademik Yaklaşımlar Dergisi (Journal of Academic Approach), 2(1), 91-113.

[2] Giray, S.,Gülel E.F.(2014). Avrupa Ülkelerinin İntihar Oranlarına Göre Sınıflandırılması, SDÜ Fen Edebiyat Fakültesi SDU Faculty of Arts and Sciences Sosyal Bilimler Dergisi Journal of Social Sciences Nisan 2014, 31, 235-247.

[3] Aşan, Z. (2007). Kredi Kartı Kullanan Müsterilerin Sosyo-Ekonomik Özelliklerinin Kümeleme Analiziyle İncelenmesi, Dumlupınar Üniversitesi Sosyal Bilimler Dergisi, Nisan 2007 (17), 256–268.

[4] Çelik Ş.(2013). Kümeleme Analizi İle Sağlık Göstergelerine Göre Türkiye'deki İllerin Sınıflandırılması, Doğuş Üniversitesi Dergisi, 14 (2) 2013, 175-194.

[5] Şişeci, M., Metlek, S., Cetişli, B.(2014) Alt-Bloklar Tekniği Ve Kümeleme Yöntemleri İle Görüntü Bölütlemenin Hızlandırılması., Gazi Üniversitesi Mühendislik-Mimarlık Fakültesi Dergisi, 29(4), 655-664.

[6] Demiralay, M., Çamurcu., A. Y. (2005). Cure, Agnes ve K-Means Algoritmalarındaki Kümeleme Yeteneklerinin Karşılaştırılması, İstanbul Ticaret Üniversitesi Fen Bilimleri Dergisi, 4(8),1-18.

[7] Özekes, S. (2003). Veri madenciliği modelleri ve uygulama alanları, İstanbul Ticaret Üniversitesi Fen Bilimleri Dergisi,3,65-82.

[8] Cengiz, D. ,Öztürk., F. (2012). Türkiye'de İllerin Eğitim Düzeylerine Göre Kümeleme Analizi İle İncelenmesi, Trakya University Journal of Social Science, 14(1), 69-84.

[9] Yavuz, Ü., Ekim, U. ve Köklü, M.(2011). Üniversite Öğrencilerin Ortak Zorunlu Derslerdeki Başarılarının K-Means Algoritması İle İncelenmesi, NWSA: Engineering Sciences, 6(1), 342-347.

[10] Sarıman, G.(2011). Veri Madenciliğinde Kümeleme Teknikleri Üzerine Bir çalışma: K-Means ve K-Medoids Kümeleme Algoritmalarının Karşılaştırılması, Journal of Natural & Applied Sciences, 15-3(2011),192-202.

[11] Silahtaroğlu, G. (2008). Veri madenciliği, Papatya yayıncılık, İstanbul, 114.

[12] Çalışkan, S. K.,Soğukpınar, İ. (2008). KxKNN: K-Means ve K En Yakın Komşu Yöntemleri İle Ağlarda Nüfuz Tespiti, EMO Yayınları, 120-124.

[13] Doğan, İ.,(2002).Kümeleme Analizi ile Seleksiyon, Turk J Vet Anim Sci, 26, 47-53.

[14] Işık, M.,Çamurcu, A. Y.,(2007) K-Means, K-Medoids Ve Bulanık C-Means Algoritmalarının Uygulamalı Olarak Performanslarının Tespiti, İstanbul Ticaret Üniversitesi Fen Bilimleri Dergisi, 11(1),31-45.