[1277]

DEĞİŞKEN VERİ PARÇALANMALARI KULLANILARAK MODEL TABANLI KÜMELEMEYE DAYALI YENİ BİR SINIFLANDIRMA YÖNTEMİ: UZAKTAN ALGILANMIŞ ÇOK BANTLI GÖRÜNTÜ SINIFLANDIRILMASI İÇİN VERİ MADENCİLİĞİ YAKLAŞIMI

Maruf GÖGEBAKAN[1], Hamza EROL[2]

[1]Arş.Gör., Abdullah Gül Üniversitesi, Uygulamalı Matematik Bölümü, 38170, Kayseri, maruf.gogebakan@agu.edu.tr
[2]Prof. Dr., Çukurova Üniversitesi, İstatistik Bölümü, 01330, Adana, herol@cu.edu.tr

ÖZET

Bu çalışmada uzaktan algılanmış çok bantlı uydu görüntü verisinde değişkenlerdeki heterojen yapıların parçalanmalarının verideki kümelenme sayısını ve yapısını belirlediği ve şekillendirdiği varsayılmış ve gösterilmiştir. Uzaktan algılanmış çok bantlı uydu görüntü verisinde değişkenlerde parçalanma olup olmadığı durum incelenmiştir. Değişkenlerdeki parçalanmalara göre veride olabilecek tüm kümelenme merkez sayıları belirlenmiştir. Tüm kümelenme merkez sayıları arasında bazıları varsayımları sağlamadığı için çıkarılmıştır. Geriye kalan kümelenme merkez sayıları mümkün kümelenme merkezlerini verir ve bunlar varsayımları sağlar. Uzaktan algılanmış çok bantlı uydu görüntü verisinde mümkün kümelenme merkezleri için kümelenme sayısını ve yapısını belirlemek amacıyla normal dağılımların karma modelleri kullanılarak aday modeller oluşturulmuştur. Oluşturulan normal dağılımların karma modelleri arasından en iyisi bilgi kriterleri kullanılarak seçilmiştir. En iyi olarak seçilen normal dağılımların karma modeli verideki kümelenme sayısını ve yapısını belirler. En iyi normal dağılımların karma modelindeki bileşen sayısı verideki kümelenme sayısına karşılık gelir. En iyi normal dağılımların karma modelindeki bileşenlerin oranları, ortalama vektörleri ve varyans – kovaryans matrisleri uzaktan algılanmış çok bantlı uydu görüntü verisindeki kümelenmelerin sırasıyla büyüklüğüne, konumuna ve yapısına karşılık gelir. Bu çalışmayla verideki heterojen değişkenlerin parçalanmasını kullanan ve normal dağılımların karma modellerinde seçime dayalı uzaktan algılanmış çok bantlı uydu görüntülerini sınıflandırmak için yeni bir veri madenciliği yöntemi geliştirilmiştir.

Anahtar Sözcükler: Veri madenciliği, Karma model kümeleme, Model seçimi, Çok bantlı görüntü verisi, Değişken veri parçalama.

ABSTRACT

A NEW CLASSIFICATION METHOD BASED ON MIXTURE MODEL CLUSTERING USING VARIABLE DATA SEGMENTATION: DATA MINING APPROACH FOR REMOTELY SENSED MULTISPECTRAL IMAGE DATA CLASSIFICATION

A new method for classification of remotely sensed multispectral image data using mixture model clustering based on model selection is proposed in this study by assuming that segmentations of each heteregeneous spectral bands forms the number of clusters and determines the structures of clusters in multispectral image data. The case of three heteregeneous spectral bands with each having segmentations is considered as real data study. The number of all cases for cluster centers determined according to the segmentations of spectral bands. Some of all cases which the assumptions are not satisfied eliminated. The rest cases gives the number of possible cluster centers which the assumptions satisfied. Candidate mixture models are established to determine the number and the structures of clusters for possible cluster centers using the partitions of heteregeneous spectral bands. The best mixture model is chosen among candidate mixture models for multispectral image data clustering using information criterions. The best mixture model thus the mixture of normal distributions determines the number and the structure of clusters in multispectral image data. The number of components in the best mixture model corresponds to the number of clusters in multispectral image data. The components of the best mixture model corresponds to the structure of clusters in multispectral image data.

Keywords: Data mining, Mixture model clustering, Model selection, Multispectral image data, Variable data segmentation.

1.GİRİŞ

Uzaktan algılanmış çok bantlı uydu görüntü verisinin normal karma kümeleme (Raftery 1998) yöntemiyle modele dayalı kümelenmesi için geliştirilen genetik algoritma ve bilgi kriterleri kullanılarak iyi kümelenmeyi veren modelin belirlenmesi sağlanmıştır. Çok bantlı uydu görüntü verisinin her bir değişkenindeki anlamlı alt gruplar modeldeki kümelenmeyi belirlemektedir (Erol 2013). Değişkendeki alt grupların oluşturduğu bütün karma modeller arasından uygun olanların seçilip uygun olmayan modellerin elenmesi ile veri kümelenme yapısına uygun aday modeller belirlenmiştir. Elde edilen aday modeller arasından en iyi modelin seçilmesi için verideki gözlemlerden parametreler elde edilmiş, bu parametre değerleri ile istatistiksel bilgi kriterleri kullanılmıştır.

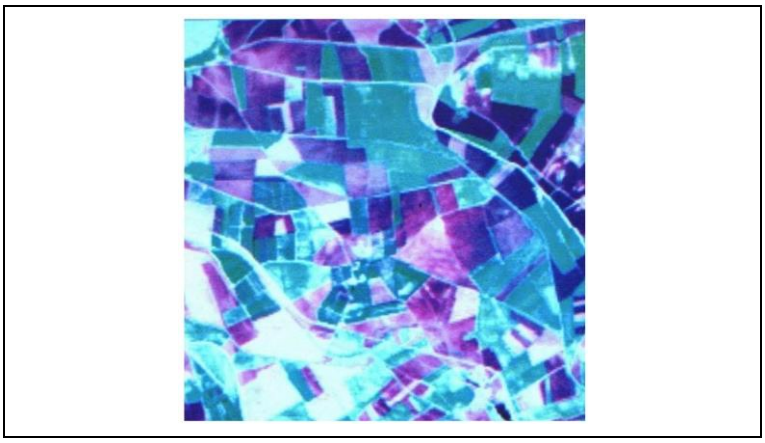
Uzaktan algılanmış çok bantlı uydu görüntü verisindeki alt alanlar modele dayalı kümeleme analizindeki anlamlı alt gruplara karşılık gelmektedir. Günümüzdeki donanımlı uydularla istenildiği kadar fazla miktarda veri toplanmakta ve elde edilen verilerin analizi için uygulanabilir, performansı bilinen diğer yöntemlere göre daha az maliyetli eğitimli ya da eğitimsiz yöntemlere gereksinim duyulmaktadır (Erol ve Akdeniz 2005).

2.YÖNTEM

Bu bölümde, çalışmada algoritma seçimi için önerilen stratejinin çalışma prensibi, WEKA veri madenciliği yazılımının: veriyi açıklama ortamı - explorer ve algoritma seçimi ortamı - experimenter ortamı - knowledge flow kullanılarak uzaktan algılanmış çok bantlı uydu görüntü verisinin sınıflandırılması uygulaması üzerinde açıklanacaktır.

2.1.Çok Bantlı Uydu Görüntü Verisi

Bu çalışmada Seyhan Ovası Adana - Türkiye (≈37^{o}N, 36^{o}E) bölgesinde yer alan tarımsal bölgenin 27 Mart 1992 tarihli (Path 175 - Row 34) 198 satır ve 200 sütundan oluşan Landsat Thematic Mapper çok bantlı uydu görüntü verisinin 3.bant, 4.bant ve 5.bant değerleri kullanılmıştır. Çalışılan tarımsal bölgenin uydu görüntüsü Şekil 1 de gösterilmiştir.



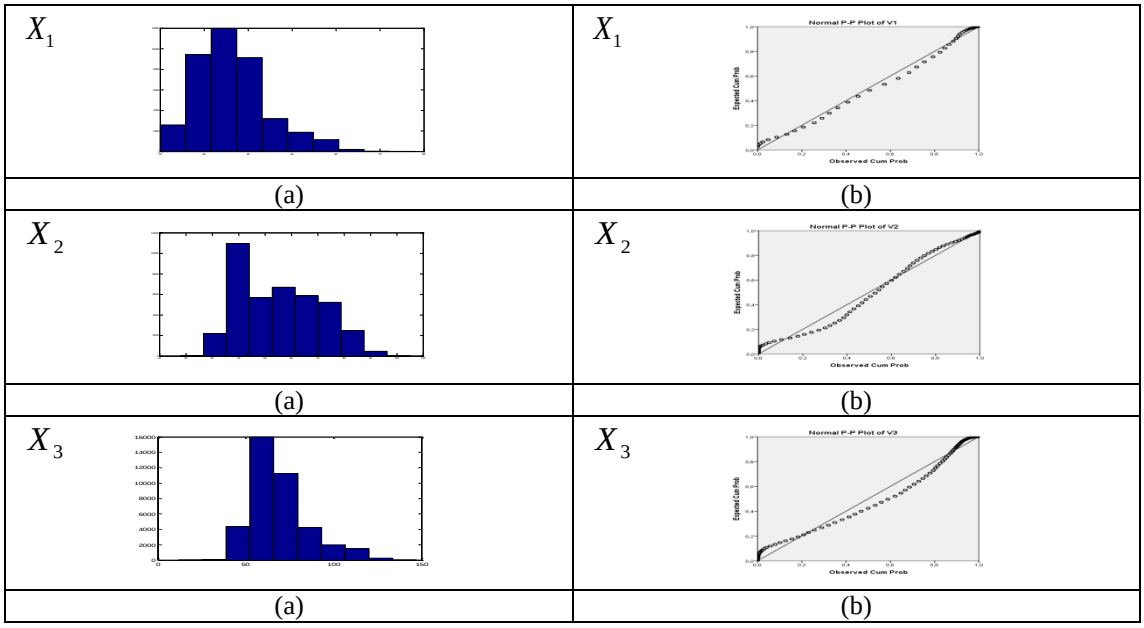
Şekil 1. Seyhan Ovası Adana - Türkiye (≈37^{o}N, 36^{o}E) bölgesinde yer alan tarımsal bölgenin 27 Mart 1992 tarihli (Path 175 - Row 34) 198 satır ve 200 sütundan oluşan Landsat Thematic Mapper uydu görüntüsü (Erol ve Akdeniz 2005).

2.2.Uydu Görüntü Verisindeki Heterojen Değişkenler İçin Uygun Parçalanma Sayılarının Belirlenmesi

Uzaktan algılanmış çok bantlı uydu görüntüsünün sayısallaştırılmış verisindeki 3.bant, 4.bant ve 5.bant değerleri bu çalışmada değişken değerleri olarak alınmıştır. Veride X_1 değişkeni 3.banttaki 198×200=39600, X_2 değişkeni 4.banttaki 198×200=39600 ve X_3 değişkeni 5.banttaki 198×200=39600 gözlem değerinden oluşmaktadır. Veride toplam 39600×3=118800 gözlem değeri bulunmaktadır. Öncelikle bu üç değişkendeki verilerin homojen mi? yani dağılımının normal mi? ya da heterojen mi? yani dağılımının normal dağılımların karması mı? olduğu araştırılır. Veri setindeki her bir bant bir değişkeni temsil ettiğinden modeller oluşturulurken bu bantlardan elde edilen değişkenlerin yapıları ya da parçalanmaları kullanılmıştır (Gögebakan ve Erol 2016). Değişkenlerdeki homojen ve heterojen yapılar, yani parçalanmalar belirlenirken hem hesaplama yöntemi hem de grafiksel yöntemler kullanılmıştır. Verideki her bir değişkenin histogramına ve P-P grafiklerine bakılarak da değişkendeki parçalanma sayısı kontrol edilerek belirlenmiştir. Çok bantlı uydu görüntü verisindeki (3.bant) X_1 değişkenindeki, (4.bant) X_2 değişkenindeki ve (5.bant) X_3 değişkenindeki parçalanmalar önce histogram ve P-P grafikleri kullanılarak grafiksel yöntemlerle incelenmiştir. Bu durum Şekil 2 de gösterilmiştir. Çok bantlı uydu görüntü verisindeki (3.bant) X_1 değişkeni, (4.bant) X_2 değişkeni ve (5.bant) X_3 değişkeni histogram ve P-P grafikleri kullanılarak grafiksel yöntemlerle inceleme sonucunda heterojen yapıda oldukları ve bu değişkenlerdeki parçalanmaların üç olduğu belirlenmiştir. Üç değişkenli uydu görüntü verisinde değişkenlerdeki parçalanma sayılarının belirlenmesi için hesaplama yönteminde her bir değişken için normal karma model kullanılarak belirlenmiştir (Erol ve Erol 2016). Normal karma model,

$$f(x)=\Sigma_{i=1}^{k} \pi_i f_i(x; \mu_i, \sigma_i) \qquad (1)$$

formundadır. Burada f(x): normal karma modelin olasılık yoğunluk fonksiyonunu; k: karma normal modeldeki bileşen (parçalanma) sayısını; π_i:bileşen ağırlığını; $f_i(x; \mu_i, \sigma_i)$: bileşenin olasılık yoğunluk fonksiyonunu; μ_i: bileşen ortalamasını; σ_i: bileşen standart sapmasını göstermektedir. Değişkendeki uygun parçalanma sayısına göre tek değişkenli normal karma modeldeki karma ağırlıkları π, ortalamalar μ ve varyanslar σ^2 tahmin edilir. Her bir değişkendeki parçalanmanın ortaya çıkarılması için tek değişkenli normal karma modelin log-likelihood, AIC ve BIC değerleri modelin parametrelerinden parçalanma sayısına bağlı olarak hesaplanır. Tek değişkenli normal karma modelden elde edilen log-likelihood değerinin en büyük, AIC ve BIC değerlerinin en küçük olduğu parçalanma sayısı en uygun parçalanma sayısı olarak belirlenir.



Şekil 2. Çok bantlı uydu görüntü verisinin (3.bant) X_1 değişkeni, (4.bant) X_2 değişkeni ve (5.bant) X_3 değişkeni için sırasıyla (a) histogram ve (b) P-P grafikleri.

Her bir değişkendeki uygun parçalanma sayısı k, bir algoritma kullanılarak elde edilebilir. Uygun parçalanma sayısı hesaplanırken algoritmada $k = 1, 2, \ldots$ değerleri verilerek elde edilen bilgi kriterlerine göre seçilir. k değerinin üst sınırı grafiksel yöntemlerden elde edilen tahmini parçalanma sayısına göre geliştirilen algoritmaya girilebilir. Ancak grafiksel yöntemlerden bağımsız olarak k için bir aralık,

$$1 \le k \le \frac{1}{k!} \sum_{k=1}^{n} \binom{n}{k} (-1)^{n-k} k^n \qquad (2)$$

şeklinde elde edilebilir (Bozdogan 1984). Burada n: değişken için gözlem sayısı ve k: değişkendeki uygun parçalanma sayısını göstermektedir. Ancak aralıktaki üst sınır verideki gözlem sayısı arttıkça üstel olarak büyüyeceği için grafiksel yöntemlerden elde edilen tahmini sayıya dayalı olarak hesaplanması kolaylık sağlamaktadır. Çok bantlı uydu görüntü verisindeki X_1 değişkeni için veri değerleri kullanılarak uygun parçalanma sayısını belirlemek amacıyla modeldeki karma ağırlıklar π, ortalamalar μ ve varyanslar σ^2 için tahmin değerlerine dayalı olarak hesaplanan log-likelihood, AIC ve BIC değerleri Çizelge 1 de verilmiştir.

Çizelge 1. X_1 için uygun parçalanma sayısını belirlemek amacıyla modeldeki karma ağırlıklar π, ortalamalar μ ve varyanslar σ^2 için tahmin değerlerine dayalı olarak hesaplanan log-likelihood, AIC ve BIC değerleri.

k	Log-L	AIC	BIC
k=1	-13947	27895	27897
k=2	-13767	27534	27538
k=3*	**-13567**	**27235**	**27242**
k=4	-13617	27236	27246
k=5	-13617	27237	27249

Çizelge 1 deki sonuçlara göre X_1 değişkeni için uygun parçalanma sayısı üçtür. Çok bantlı uydu görüntü

verisindeki X_2 değişkeni için veri değerleri kullanılarak uygun parçalanma sayısını belirlemek amacıyla modeldeki karma ağırlıklar π, ortalamalar μ ve varyanslar σ^2 için tahmin değerlerine dayalı olarak hesaplanan log-likelihood, AIC ve BIC değerleri Çizelge 2 de verilmiştir.

Çizelge 2. X_2 için uygun parçalanma sayısını belirlemek amacıyla modeldeki karma ağırlıklar π, ortalamalar μ ve varyanslar σ^2 için tahmin değerlerine dayalı olarak hesaplanan log-likelihood, AIC ve BIC değerleri.

k	Log-L	AIC	BIC
k=1	-16418	32837	32838
k=2	-15912	31825	31829
k=3*	**-15858**	**31717**	**31724**
k=4	-15857	31715	31725
k=5	-15857	31716	31728

Çizelge 2 deki sonuçlara göre X_2 değişkeni için uygun parçalanma sayısı üçtür. Çok bantlı uydu görüntü verisindeki X_3 değişkeni için veri değerleri kullanılarak uygun parçalanma sayısını belirlemek amacıyla modeldeki karma ağırlıklar π, ortalamalar μ ve varyanslar σ^2 için tahmin değerlerine dayalı olarak hesaplanan log-likelihood, AIC ve BIC değerleri Çizelge 3 te verilmiştir.

Çizelge 3. X_3 için uygun parçalanma sayısını belirlemek amacıyla modeldeki karma ağırlıklar π, ortalamalar μ ve varyanslar σ^2 için tahmin değerlerine dayalı olarak hesaplanan log-likelihood, AIC ve BIC değerleri.

k	Log-L	AIC	BIC
k=1	-16664	33329	33331
k=2	-16169	32340	32344
k=3*	**-16126**	**32254**	**32261**
k=4	-16134	32330	32339
k=5	-16138	32340	32352

Çizelge 3 teki sonuçlara göre X_3 değişkeni için uygun parçalanma sayısı üçtür. Uzaktan algılanmış çok bantlı uydu görüntü verisindeki X_1, X_2 ve X_3 değişkenleri için log-likelihood fonksiyon değerinin maksimumu, aynı zamanda AIC ve BIC değerlerinin minimumu kullanılarak uygun parçalanma sayısını veren karma normal model: X_1 değişkeni için k=3, X_2 değişkeni için k=3 ve X_3 değişkeni için k=3 olarak belirlenmiştir.

2.3.Çok Bantlı Uydu Görüntü Verisinde Değişkendeki Uygun Parçalanmaların Belirlenmesi

Üç değişkenli ve değişkenlerin her birinin üçe parçalandığı veri setindeki değişkenlerin anlamlı alt gruplarına düşen gözlem değerlerinin belirlenmesi için k-ortalamalar algoritması kullanılmıştır. Başlangıçta her bir değişkenin parçalanma sayısı kadar k merkez sayısı belirlenerek adımsal işlemlerle gözlemler arasındaki uzaklıklara göre merkez etrafındaki en yakın gözlemler parçalanmalara atanmaktadır. Seçilen giriş küme merkezi değeri ile gözlemler arasındaki uzaklık,

$$\arg\min_{s} \sum_{i=1}^{k} \sum_{x_j \in S_i} \left\| x_j - \mu_i \right\|^2 \qquad (3)$$

şeklinde hesaplanmaktadır. Burada her bir grup ya da parçalanma için gözlem değeri ile grup merkezi arasındaki uzaklıkların toplamı alınarak her bir gözlem değeri için en uygun grup merkezi seçilir. Üç değişkenli ve değişkenlerin her birinde üç anlamlı alt grup bulunan veride k-ortalamalar algoritması uygulandığında her bir değişken için üç ayrı küme merkezi belirlenmiş ve her bir küme merkezinin etrafına aralarındaki mesafe minimum olacak şekilde gözlem değerleri atanmıştır. Kümeler arası mesafenin maksimum (heterojen) aynı zamanda küme içi mesafenin minimum (homojen) olduğu durum en uygun parçalanma sayısını verir. K-ortalamalar algoritması kullanılarak verideki her bir değişken üçe parçalanmış ve her bir gözlemin düştüğü anlamlı alt grup belirlenmiştir. Üç değişkenli veride k-ortalamalar algoritması uygulandığında X_1, X_2 ve X_3 değişkenleri için parçalanma

sayıları sırasıyla $k_1 = 3$, $k_2 = 3$ ve $k_3 = 3$ olarak elde edilmiştir. X_1 değişkenindeki parçalanmalar: X_{11}, X_{12} ve X_{13}; X_2 değişkenindeki parçalanmalar: X_{21}, X_{22} ve X_{23}; X_3 değişkeninde parçalanmalar: X_{31}, X_{32} ve X_{33} olarak tanımlanır ve alınır. Değişkenlerdeki parçalanmaların gözlem sayıları Çizelge 4 de verilmiştir.

Çizelge 4. Üç değişkenli uydu görüntü verisinde değişkenlerdeki parçalanmalara düşen gözlem sayıları.

Değişken	X_1			X_2			X_3		
Değişken parçalanmaları	X_{11}	X_{21}	X_{31}	X_{21}	X_{22}	X_{23}	X_{31}	X_{32}	X_{33}
Parçalanmadaki Gözlem sayısı	4291	17241	18068	10279	12697	16624	17929	4999	16672
Toplam	39600			39600			39600		

2.4.Çok Bantlı Uydu Görüntü Verisinde Değişkenlerdeki Uygun Parçalanmalara Dayalı Kümelenme Merkez Sayılarının Ve Yapılarının Belirlenmesi

Uzaktan algılanmış çok bantlı uydu görüntü verisindeki X_1, X_2 ve X_3 değişkenlerindeki parçalanmalar sırasıyla $k_1 = 3$, $k_2 = 3$ ve $k_3 = 3$ olarak elde edilmiştir. Çok değişkenli verideki değişken sayısı $p = 3$ alınmıştır. Üç değişkenli veri seti için C_{min} ve C_{mak} ile gösterilen değişkenlerdeki parçalanmaların oluşturduğu kümelenme merkezlerinin sırasıyla minimum ve maksimum sayısı, Servi ve Erol (2007) tarafından önerilen

$C_{min} = k_1 k_2 k_3$

ve (4)

$$C_{mak} = \prod_{s=1}^{p} k_s$$

eşitliğiyle hesaplanır. Burada p, verideki değişken sayısını ve k_s, X_s değişkenindeki parçalanma sayısını göstermektedir. $n \times 3$ tipindeki X veri matrisi $X = \lfloor X_1 \quad X_2 \quad X_3 \rfloor$ şeklinde matris formunda gösterilebilir. n_1 elemanlı X_1 değişkenindeki parçalanma $X_1 = \begin{bmatrix} X_{11} \\ X_{12} \\ X_{13} \end{bmatrix}$ şeklinde olur. Burada X_{11}, X_{12} ve X_{13} sırasıyla n_{11}, n_{21} ve n_{13} elemanlıdır. Yani $n_1 = n_{11} + n_{12} + n_{13}$ dir. Burada n_1, X_1 değişkenindeki eleman sayısıdır. n_2 elemanlı X_2 değişkenindeki parçalanma $X_2 = \begin{bmatrix} X_{21} \\ X_{22} \\ X_{23} \end{bmatrix}$ şeklinde olur. Burada X_{21}, X_{22} ve X_{23} sırasıyla n_{21}, n_{22} ve n_{23} elemanlıdır. Yani $n_2 = n_{21} + n_{22} + n_{23}$ dir. Burada n_2, X_2 değişkenindeki eleman sayısıdır. n_3 elemanlı X_3 değişkenindeki parçalanma $X_3 = \begin{bmatrix} X_{31} \\ X_{32} \\ X_{33} \end{bmatrix}$ şeklinde olur. Burada X_{31}, X_{32} ve X_{33} sırasıyla n_{31}, n_{32} ve n_{33} elemanlıdır. Yani $n_3 = n_{31} + n_{32} + n_{33}$ dir. Burada n_3, X_3 değişkenindeki eleman sayısıdır. Bu durumda (4) deki eşitlik kullanılarak üç değişkendeki parçalanmaların oluşturduğu kümelenme merkezlerinin minimum ve maksimum sayısı, $C_{\min} = \max\{k_1, k_2, k_3\} = \max\{3,3,3\} = 3$ ve $C_{mak} = k_1 k_2 k_3 = 3.3.3 = 27$ olarak bulunur. Üç değişkenli uzaktan algılanmış çok bantlı uydu görüntü verisinde X_1, X_2 ve X_3 değişkenlerindeki sırasıyla

X_{11}, X_{12} ve X_{13}; X_{21}, X_{22} ve X_{23}; X_{31}, X_{32} ve X_{33} parçalanmalara karşılık gelen C_{mak} ile hesaplanan 27 kümelenme merkezi ve bu merkezleri meydana getiren değişkenlerdeki bileşenler Çizelge 5 te verilmiştir.

Çizelge 5. Üç değişkenli uzaktan algılanmış çok bantlı uydu görüntü verisinde X_1, X_2 ve X_3 değişkenlerindeki sırasıyla X_{11}, X_{12} ve X_{13}; X_{21}, X_{22} ve X_{23}; X_{31}, X_{32} ve X_{33} parçalanmalara karşılık gelen C_{mak} ile hesaplanan 27 kümelenme merkezi ve bu merkezleri meydana getiren değişkenlerdeki bileşenler.

Merkez No	Merkez Bileşenleri	Merkez No	Merkez Bileşenleri	Merkez No	Merkez Bileşenleri
1.	(X_{11}, X_{21}, X_{31})	10.	(X_{11}, X_{21}, X_{32})	19.	(X_{11}, X_{21}, X_{33})
2.	(X_{12}, X_{21}, X_{31})	11.	(X_{12}, X_{21}, X_{32})	20.	(X_{12}, X_{21}, X_{33})
3.	(X_{13}, X_{21}, X_{31})	12.	(X_{13}, X_{21}, X_{32})	21.	(X_{13}, X_{21}, X_{33})
4.	(X_{11}, X_{22}, X_{31})	13.	(X_{11}, X_{22}, X_{32})	22.	(X_{11}, X_{22}, X_{33})
5.	(X_{12}, X_{22}, X_{31})	14.	(X_{12}, X_{22}, X_{32})	23.	(X_{12}, X_{22}, X_{33})
6.	(X_{13}, X_{22}, X_{31})	15.	(X_{13}, X_{22}, X_{32})	24.	(X_{13}, X_{22}, X_{33})
7.	(X_{11}, X_{23}, X_{31})	16.	(X_{11}, X_{23}, X_{32})	25.	(X_{11}, X_{23}, X_{33})
8.	(X_{12}, X_{23}, X_{31})	17.	(X_{12}, X_{23}, X_{32})	26.	(X_{12}, X_{23}, X_{33})
9.	(X_{13}, X_{23}, X_{31})	18.	(X_{13}, X_{23}, X_{32})	27.	(X_{13}, X_{23}, X_{33})

Üç değişkenli normal dağılımların karma modelleri kullanılarak elde edilen modele dayalı kümelemede bu merkezler ve merkezlerden elde edilen modeller araştırılacaktır. Her bir merkezi meydana getiren değişkenlerin parçalanmaları, modellerin hesaplamalarında kullanılacak parametrelerin elde edilmesinde kullanılmıştır.

2.5.Çok Bantlı Uydu Görüntü Verisinde Toplam Model Sayısı Ve Modellerin Yapısının Belirlenmesi

Uzaktan algılanmış çok bantlı uydu görüntü verisindeki X_1, X_2 ve X_3 değişkenlerindeki parçalanmaların oluşturduğu kümelenme merkezleri için M_{Toplam} ile gösterilen üç değişkenli normal dağılımların karma modelleriyle oluşturulabilecek toplam model sayısı,

$$M_{Toplam} = 2^{C_{mak}} - 1 \qquad (5)$$

eşitliğinden elde edilir (Servi ve Erol 2007). Uzaktan algılanmış çok bantlı uydu görüntü verisi için M_{Toplam} değeri, $M_{Toplam} = 2^{3.3.3} - 1 = 2^{27} - 1 = 134.217.727$ olarak elde edilir. Burada çıkarılan 1 model, sabit modeldir.

2.6.Çok Bantlı Uydu Görüntü Verisinde Değişkenlerdeki Parçalanmalara Dayalı Uygun Aday Model Sayısının Hesaplanması

Üç değişkenli uzaktan algılanmış çok bantlı uydu görüntü verisi için normal karma modellerde değişkenlerdeki parçalanmalara karşılık gelen kümelenme merkezleri ve parçalanmalara bağlı olarak toplam model sayısı elde edilmiştir. Normal karma modeller oluşturulurken değişkenlerdeki her parçalanmaya en az bir kümelenme merkezi karşılık gelecek şekildeki geçerli veya uygun modellerin sayısı araştırılmıştır. Bu aday modeller Çizelge 5 de belirtilen 27 merkez üzerinden her boyutta en az bir kümelenme bulunacak varsayımı ile elde edilmiştir. Çizelge 5 deki değişkenlerin parçalanmalarının oluşturduğu kümelenme merkezleri C_{min} ve C_{mak} eşitliklerinden elde edilen en az 3 ve en fazla 27 kümelenmenin olduğu varsayımı yapılmıştır. Varsayıma uyan modellere uygun aday model denilmektedir. Uygun aday model sayısı değişken sayısı ve değişkenlerdeki parçalanma sayısına bağlı

olarak hesaplanmıştır. Her boyutta (bantta) en az bir merkezin bulunduğu varsayımına uyan modellerin sayısı, her bir merkezin bulunduğu boyuttaki konumuna göre kombinasyon hesabı ile elde edilebilir. Uzaktan algılanmış çok bantlı uydu görüntü verisindeki değişken sayısı ve değişkendeki gözlem sayısı büyük veri olduğundan hesaplama geliştirilen algoritma ile elde edilmiştir. Uzaktan algılanmış çok bantlı uydu görüntü verisindeki her bir değişken ve değişkenlerdeki parçalanmalara dayalı aday modellerin sayısı hesaplanmıştır. Değişkenlerin parçalanmasından sonra oluşan merkez sayısı, toplam model sayısı ve uygun aday model sayısı Çizelge 6 da verilmiştir.

Çizelge 6. Çok bantlı uydu görüntü verisindeki değişkenlerin parçalanmalarına dayalı oluşan merkez sayısı, toplam model sayısı ve uygun aday model sayısı.

Merkez Sayısı	Toplam Model Sayısı	Uygun Aday Model Sayısı
1	27	0
2	351	0
3	2925	36
4	17550	1890
5	80730	24300
6	296010	153828
7	888030	623106
8	2220075	1839672
9	4686825	4255194
10	8436285	8044245
11	13037895	12751803
12	17383860	17216811
13	20058300	19981143
14	20058300	20030760
15	17383860	17376516
16	13037895	13036518
18	8436285	8436123
18	4686825	4686816
19	2220075	2220075
20	888030	888030
21	296010	296010
22	80730	80730
23	17550	17550
24	2925	2925
25	351	351
26	27	27
27	1	1
Toplam	134217728	131964460

Uygun aday modellerin sayısı Çizelge 6 da bilgisayar bilimleri yöntemiyle hesaplandığı gibi aşağıdaki denklem yardımıyla da elde edilir (Cheballah ve ark. 2015).

$$f(n,m,s,k) = \sum_{i=0}^{n} (-1)^i \binom{n}{i} \sum_{j=0}^{m} (-1)^j \binom{m}{j} \sum_{t=0}^{s} (-1)^t \binom{s}{t} \binom{(n-i)(m-j)(s-t)}{k}$$

$$= \sum_{i=0}^{n} \sum_{j=0}^{m} \sum_{t=0}^{s} (-1)^{i+j+t} \binom{n}{i}\binom{m}{j}\binom{s}{t} \binom{(n-i)(m-j)(s-t)}{k}$$

$$= \sum_{i,j,t=0}^{n,m,s} (-1)^{i+j+t} \binom{n}{i}\binom{m}{j}\binom{s}{t} \binom{(n-i) \quad (m-j) \quad s-t}{k} \qquad (6)$$

Burada n, m ve s sırasıyla değişkenlerdeki parçalanma sayılarını göstermektedir. i, j ve t sırasıyla değişkenlerdeki parçalanmalardan kaynaklanan kümelenme merkez sayılarını ve k modeldeki kümelenme merkez sayısını göstermektedir.

2.7 2.7.Çok Bantlı Uydu Görüntü Verisindeki Değişkenlerin Parçalanmalarına Dayalı Uygun Durumlar İçin Aday Normal Karma Modellerin Oluşturulması

Çok bantlı uydu görüntü verisindeki değişkenlerdeki parçalanmalardan oluşan kümelenmeler için uygun durumlarda ortaya çıkan aday normal karma modellerin bileşen ağırlıkları, ortalama vektörleri ve varyans-

kovaryans matrisleri verideki heterojen değişkenlerdeki parçalanmalar kullanılarak örnekleme dayalı tahmin edilmektedir (Erol 2012). Heterojen değişkenlerin parçalanmalarıyla oluşan her bir kümelenme merkezi için aday normal karma modellerin bileşenlerinin karma ağırlıklarının, ortalama vektörlerinin ve varyans-kovaryans matrislerinin tahminleri Çizelge 6 daki uygun aday modellerin karma olasılık yoğunluk fonksiyonlarını elde etmek için kullanılır. Çizelge 6 da uygun aday modeller için merkez sayıları kullanılarak, 0 ve/veya 1 lerden oluşan ve 27 karakterden oluşan string gösterimleri yapılır. Bu string gösterimlerinde modeli oluşturan merkeze karşılık olarak "1", modeli oluşturmayan merkeze karşılık olarak "0" yazılır. Uygun modeller için temsili string gösterimler, uygun modeller için hesaplamalarda kullanılmıştır.

2.8.Çok Bantlı Uydu Görüntü Verisindeki Değişkenlerin Parçalanmalarına Dayalı Uygun Durumlar İçin Aday Normal Karma Modellerdeki Parametrelerin Tahmini

Veride değişkenlerdeki parçalanmalara karşılık gelen merkezlerdeki kümelenmeler için karma oranları, ortalama vektörleri ve varyans-kovaryans matrisleri, örneklemden tahmin edilmiştir. Üç değişkenli veride i=1,2,...,27 olmak üzere ortalama vektörleri ve varyans-kovaryans matrisleri sırasıyla, $\mu_i = \begin{bmatrix} \mu_{1\otimes} \\ \mu_{2\bullet} \\ \mu_{3*} \end{bmatrix}$ ve

$$\Sigma_i = \begin{bmatrix} \sigma_{1\otimes}{}^2 & \rho_i\sigma_{1\otimes}\sigma_{2\bullet} & \rho_j\sigma_{1\otimes}\sigma_{3*} \\ \rho_i\sigma_{2\bullet}\sigma_{1\otimes} & \sigma_{2\bullet}{}^2 & \rho_k\sigma_{2\bullet}\sigma_{3*} \\ \rho_j\sigma_{3*}\sigma_{1\otimes} & \rho_k\sigma_{3*}\sigma_{2\bullet} & \sigma_{3*}{}^2 \end{bmatrix}$$ ile temsil edilir. Burada $\otimes: X_1$ değişkenindeki 1,2,3 parçalanmayı, $\bullet: X_2$ değişkenindeki 1,2,3 parçalanmayı, $*: X_3$ değişkenindeki 1,2,3 parçalanmayı göstermektedir. $\rho_i = \boldsymbol{Corr(X}_{1\otimes}, \boldsymbol{X}_{2\bullet}, \boldsymbol{X}_{3*})$ Pearson korelasyon katsayılarını göstermektedir. Herbir kümelenme merkezindeki veriler, $N(\mathbf{x}; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ normal dağılımına sahip olsun. Üç değişkenli veride değişkenlerdeki parçalanmalara karşılık gelen kümelenme merkezleri ile normal karma modeller oluşturulurken karma ağırlıkları, ortalama vektörleri ve varyans-kovaryans matrisleri kullanılır. Değişkenlerdeki parçalanmalara karşılık gelen kümelenme merkezlerinin varyans-kovaryans matrislerindeki bileşenler arasındaki ilişkinin yönünü ve derecesini veren Pearson korelasyon katsayısı üç bileşenli merkez yapısı için $3x3$ tipinde matris içindeki bileşenlerde altı adet bulunmaktadır. Varyans-kovaryans matrisi simetrik matris yapısında olduğundan her bir varyans-kovaryans matrisinde üç farklı Pearson korelasyon katsayısı bulunur. Modellerdeki farklı korelasyon katsayılarının sayısı değişkenlerdeki parçalanmalar kullanılarak

$$k_1k_2 + k_1k_3 + k_2k_3 = 3.3 + 3.3 + 3.3 = 27 \quad (7)$$

şeklinde elde edilir.

2.9.Çok Bantlı Uydu Görüntü Verisindeki Değişkenlerin Parçalanmalarına Dayalı Uygun Durumlar İçin Aday Normal Karma Modellerin Log-Likelihood Fonksiyonu, AIC ve BIC değerleri

Heterojen veride en iyi kümelenme yapısını normal dağılımların karma modellerini kullanarak modele dayalı belirlemek amacıyla her uygun model için birinci kriter olarak log-likelihood fonksiyonu değeri hesaplanır. Uygun modeller için log-likelihood fonksiyonların değerleri en iyi modeli seçmek için bir kriter olarak kullanılmıştır. Çok değişkenli normal karma modellerin log-likelihood fonksiyonu, $i = 1, 2, ...n \;\; ve \;\; j = 1, 2, ...g$ olmak üzere $f\left(x_i; \theta_j\right)$ karma olasılık yoğunluk fonksiyonu olmak üzere,

$$L(\Psi) = \prod_{i=1}^{n} f\left(x_i; \Psi\right) = \prod_{i=1}^{n} \left[\sum_{j=1}^{g} f\left(x_i; \theta_j\right) \right] \quad (8)$$

şeklinde elde edilir. Logaritması alınmış likelihood fonksiyonu,

$$\log L(\Psi) = \sum_{j=1}^{n} \log\left(f\left(x_i; \Psi\right)\right) = \sum_{j=1}^{n} \log\left(\sum_{i=1}^{g} \pi_i f_i\left(x_j; \theta_i\right)\right) \quad (9)$$

olarak elde edilir. Cezalı (adjusted) log-likelihood fonksiyonu $\log L(\Psi)$ fonksiyonuna ceza terimi eklenerek elde edilir. Veri setinde parametreler EM algoritması gibi adımsal işlemler veya veriyi standartlaştırma gibi işlemler kullanılmadan gözlem değerlerinden elde edildiği için aşırı değerlerden veya gürültü verilerinden (noise data) etkilenmektedir. Ceza terimi eklemedeki amaç verideki aşırı değerleri veya diğer bir ifade ile gürültü verilerinin fonksiyona kattığı fazla kümelenmeyi engellemektir Schwarz (1978). Düzeltilmiş log-likelihood olarak adlandırılan (adjlog-likelihood) fonksiyon değeri,

$$adj\left(\log L(\Psi)\right) = \sum_{j=1}^{n} \log\left(\sum_{i=1}^{g} \pi_i f_i\left(x_j;\theta_i\right)\right) - d\log n \tag{10}$$

olarak elde edilir. Burada n olasılık yoğunluk fonksiyonundaki gözlem sayısını, d modeldeki bağımsız parametre sayısını göstermektedir. Çok değişkenli normal karma modellerin adj(log-likelihood) fonksiyonuna bağlı olarak Bayesci bilgi kritri (BIC),

$$BIC = -2\ln L(\Psi) + d\log n \tag{11}$$

olarak elde edilir. Burada d modeldeki bağımsız parametre sayısı, n yoğunluk fonksiyonundaki gözlem sayısı, K modeldeki bileşen sayısı ve p değişken sayısı olamak üzere,

$$d = (K-1) + (Kp) + \left(Kp\frac{(p+1)}{2}\right) \tag{12}$$

olarak elde edilir. Aynı şekilde Akaike bilgi kriteride (AIC) adj(log-likelihood) fonksiyonuna bağlı olarak,

$$AIC = -2\ln L(\Psi) + 2d \tag{13}$$

şeklinde elde edilir.

Çizelge 7. Çok bantlı uydu görüntü verisinde normal karma dağılımların modele dayalı kümelenmesi için adj(log-l), AIC ve BIC değerleri.

***adj*(Log-l)**	***adj*(AIC)**	***adj*(BIC)**	**Matris gösterimi**
-105752585,4	211505668,8433	211507806,9028	1101111111111101111111111111

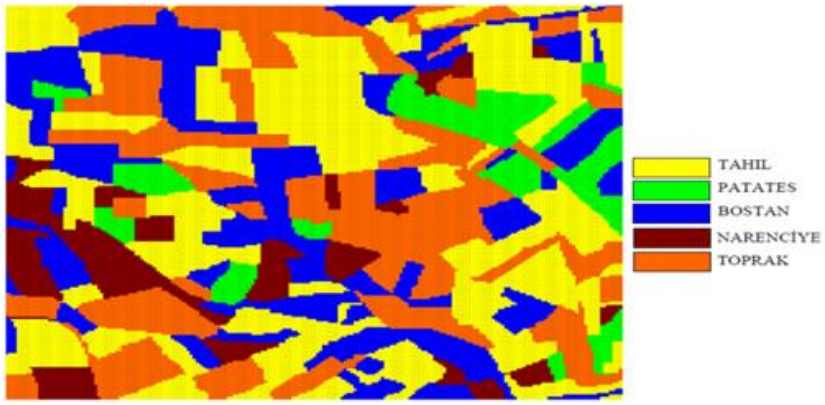
2.10.Çok Bantlı Uydu Görüntü Verisinin Normal Karma Modellerin Modele Dayalı Kümelenmesi İçin En İyi Modelin Seçimi

Modele dayalı kümeleme için çok değişkenli normal karma modeller arasından en iyi modelin seçimi modellerin adj(log-l), AIC ve BIC değerlerine dayalı olarak belirlenmektedir. Üç değişkenli veride normal karma modellerden uygun aday modellerin adj(log-l), AIC ve BIC değerleri büyük veriden algoritma yarımı ile 131964460 uygun model arasından hesaplanmış ve Çizelge 7 de verilmiştir.

Uygun aday modeller arasından modele dayalı kümelemede en iyi model adj(likelihood) değer en büyük aynı zamanda AIC ve BIC değerleri en küçük olan modeldir. Uygun aday model modelin tamamının log-likelihood, AIC ve BIC değerleri hesaplanmış aralarından en iyi model yirmi beş merkezli modeller arasından 60. sıradaki“1101111111111101111111111111” temsili gösterime sahip model olarak belirlenmiştir.

3.SONUÇLAR

Verideki heterojen değişkenlerin parçalanmaları kullanılarak uzaktan algılanmış çok bantlı uydu görüntü verisinin karma normal modele dayalı kümelenmesinden sonra oluşan herbir kümenin içeriği eğitimli sınıflandırma yöntemiyle gerçekleştirilmiştir. Eğitimli sınıflandırmada Çalış ve Erol (2013) tarafından çalışılan kontrol sınıfları kullanılmıştır. Sınıflandırma sonucunda %95 lik bir sınıflandırma doğruluk yüzdesi elde edilmiştir. Uzaktan algılanmış çok bantlı uydu görüntü verisinin kümelenmesinden sonra sınıflandırma sonucu renklendirilmiş haritası ve her rengin temsil ettiği kategoriler Şekil 3 te gösterilmiştir.



Şekil 3. Uzaktan algılanmış çok bantlı uydu görüntü verisinin kümelenmesinden sonra sınıflandırma sonucu renklendirilmiş haritası ve her rengin temsil ettiği kategoriler.

KAYNAKLAR

Akaike, H., 1974. A new look at the statistical model identification. IEEE Transactions on Automatic Control 19 (6): 716–723.

Bouveyron, C., and Brunet-Saumard, C., 2014. Model-based clustering of high-dimensional data: A review. Computational Statistics & Data Analysis, 71, 52-78.

Bozdogan, H., 1984. Multi-Sample Cluster Analysis As An Alternative To Multiple Comparison Procedures (No. UIC/DQM/A84-3). ILLINOIS UNIV AT CHICAGO CIRCLE DEPT OF QUANTITATIVE METHODS.

Cheballah, H., Giraudo, S., & Maurice, R., 2015. Hopf algebra structure on packed square matrices. Journal of Combinatorial Theory, Series A, 133, 139-182.

Çalış, N. and Erol, H., 2013. A new per-field classification method using mixture discriminant analysis. Journal of Applied Statistics 39(10):1-12. DOI: 10.1080/02664763.2012.702263.

Erol, H., 2013. A model selection algorithm for mixture model clustering of heterogeneous multivariate data. In Innovations in Intelligent Systems and Applications. 2013 IEEE International Symposium on Innovations in Intelligent Systems and Applications, At Albena, Bulgaria. (pp. 1-7). DOI: 10.1109/INISTA.2013.6577617

Erol, H. ve Akdeniz, F., 2005. A per-field classification method based on mixture distribution models and an application to Landsat Thematic Mapper data. International Journal of Remote Sensing 26(6):1229-1244.

Erol, H. and Erol, R., 2016. Logical Circuit Design Using Orientations Of Clusters In Multivariate Data For Decision Making Predictions: A Data Mining And Artificial Intelligence Algorithm Approach. International Symposium on INnovations in Intelligent SysTems and Applications 2-5 August 2016, At Sinaia, Romania, Volume: 1.

Fraley, C. and Raftery, A. E., 1998. How Many Clusters? Which Clustering Method? Answers via Model-Based Cluster Analysis. The Computer Journal, 41, 578-588.

Fraley, C. and Raftery, A. E., 2002. Model-Based Clustering, Discriminant Analysis, and Density Estimation. Journal of the American Statistical Association, 97, 611-631.

Gögebakan, M. and Erol, H., 2016. A New Approach For Mixture Model Clustering Based On Selecting The Best Mixture Model Among Candidate Mixture Models. International Conference on Information Complexity and Statistical Modeling in High Dimensions with Applications, At Nevşehir Turkey, Volume: 1.

Gögebakan, M. and Erol, H., 2016. Mixture model clustering using variable data segmentation. Conference: 12th German Probability and Statistics Days, At Bochum, Germany.

Schwarz, G., 1978. Estimating the dimension of a model, Ann. Statist. 6 pp. 461–464.

Servi, T. and Erol, H., 2007. On Total Number Of Candidate Component Cluster Centers And Total Number of Candidate Mixture Models In Model Based Clustering. Selçuk Journal of Applied Mathematics Vol.8. No.2. pp. 57 – 69.