NÖHÜ Müh. Bilim. Derg. / NOHU J. Eng. Sci., Erken Görünüm / InPress
Niğde Ömer Halisdemir Üniversitesi Mühendislik Bilimleri Dergisi
Niğde Ömer Halisdemir University Journal of Engineering Sciences
Araştırma makalesi / Research article
www.dergipark.org.tr/tr/pub/ngumuh / www.dergipark.org.tr/en/pub/ngumuh

Hiperspektral görüntülerde Relief-F algoritması ile band seçimi

Band selection with Relief-F algorithm in hyperspectral ımages

Mehmet Yılmaz[1,*], Ümit Haluk Atasever[2]

[1] *Kayseri Üniversitesi, Mimarlık ve Şehir Planlama Bölümü, 38900, Kayseri Türkiye*
[2] *Erciyes Üniversitesi, Harita Mühendisliği Bölümü, 38030, Kayseri Türkiye*

Öz

Hiperspektral görüntüler, sınıflandırması için detaylı bilgi içermektedirler. Ancak bu veriler yüksek boyut, büyük veri hacmi ve bitişik bantlar arasındaki güçlü korelasyon özellikleri nedeniyle sınıflandırma sonuçlarını olumsuz etkilenmektedir. Uygun bir öznitelik seçim yöntemi ile hiperspektral görüntülerin sınıflandırma etkinliği ve doğruluğu iyileştirilebilir. Bu çalışmada sınıflandırma modelinden bağımsız olması, çoklu bağlantı varsayımını dikkate almaması, gürültü değerlerini işleyebilmesi gibi özellikleri nedeniyle Relief-F öznitelik seçme algoritması tercih edilmiştir. Relief-F algoritmasının uygulama etkisini incelemek için Salinas-A, Indian Pines ve Pavia University veri setleri, deneysel veri olarak kullanılmıştır. Gerçekleştirilen uygulamalar sonrasında band seçimi sonrası Salinas-A, Indian Pines verisetlerinde Destek Vektör Makineleri sınıflandırıcısının daha yüksek performans gösterirken; Rastgele Orman yönteminin sınıflandırma doğruluğunun büyük oranda korunduğu görülmüştür. Araştırma sonuçları, Relief-F algoritmasının hiperspektral görüntülerde en gerekli özelliklerini belirlemek ve iyi bir sınıflandırma doğruluğu ile bant sayısının %60 - %70 azaltılabileceği göstermektedir.

Anahtar kelimeler: Hiperspektral görüntü, Sınıflandırma, Öznitelik seçimi, Relief-F

Abstract

Hyperspectral images contain detailed information for classification. However, these data negatively affect the classification results due to their high size, large data volume and strong correlation between adjacent bands. Classification efficiency and accuracy of hyperspectral images can be improved with an appropriate feature selection method. In this study, the Relief-F feature selection algorithm was preferred due to its features such as being independent of the classification model, not taking into account the assumption of multicollinearity, and being able to process noise values. Salinas-A, Indian Pines and Pavia University datasets were used as experimental data to examine the application effect of the Relief-F algorithm. After the applications, the Support Vector Machine classifier showed higher performance in the Salinas-A and Indian Pines datasets after band selection; It has been observed that the classification accuracy of the Random Forest method is largely preserved. The research results show that the Relief-F algorithm determines the most necessary features in hyperspectral images and the number of bands can be reduced by 60% - 70% with a good classification accuracy.

Keywords: Hyperspectral image, Classification, Feature selection, Relief-F

1 Giriş

Hiperspektral görüntüler, görünürden kızılötesi bölgelere kadar değişen yüzlerce dar ve bitişik elektromanyetik spektral bantlardan (görüntülerden) oluşur. Hiperspektral görüntüler arazi örtüsü sınıflarını tanımlamak ve ayırt etmek için yeterli spektral bilgiye sahiptir. Bir hiperspektral görüntünün her pikseli, bant sayısı uzunluğunda bir vektör ile temsil edilir [1]. Hiperspektral görüntülerde sınıflandırma süreçlerini zorlaştıracak sayıda band içerebilir [2]. Bu bantlar bir yandan hesaplama karmaşıklığını arttırırken diğer yandan sınıflandırma doğruluğunu azaltır [3]. Hiperspektral görüntülerde Destek Vektör Makinesi (DVM), K-En yakın komşuluk (k-EYK) gibi sınıflandırma yöntemleri öznitelik seçimi yapılmadan doğrudan kullanıldığında sınıflandırma doğruluğu genellikle düşüktür. Ayrıca tüm bantların doğrudan kullanılması, hiperspektral görüntülerin sınıflandırılmasında önemli sorunlara neden olur [4].

Hiperspektral verilerin daha büyük veri boyutu, daha fazla sayıda sınıfın elde edilmesini sağlar. Ancak hiperspektral görüntüdeki öznitelik sayısı artıkça, arazi örtüsü sınıfları düzgün bir şekilde sınıflandırmak için, örneklem büyüklüğünün de artması gerekir. Bu durum, Hughes fenomeni veya boyutsallığın laneti olarak bilinir [1]. Hiperspektral görüntülerin sınıflandırılmasında her sınıf için eşit sayıda eğitim örneği varsa istatistiksel bütünlük korunabilir. İstatistiksel sınıflandırma algoritmalarında sınıflara özgü spektral yanıt değerlerinin varyansını ve kovaryansını güvenilir bir şekilde değerlendirmek için her bir eğitim sınıfındaki piksel sayısının 10N ile 100N arasında olması tercih edilir. Burada N spektral band sayısıdır [5].

Hiperspektral görüntülerle ilgili bir başka sorunda, komşu bantların genellikle güçlü bir şekilde ilişkili olmasıdır. Bu nedenle hiperspektral görüntülerde spektral çözünürlük artırılsa bile hiçbir yeni bilgi eklenmeyebilir [1]. Hiperspektral görüntülerin yüksek boyutluluğu ve eğitim verilerinin sınırlı sayısı olması, denetimli sınıflandırma yaklaşımlarının doğru istatiksel tahminler yapabilmesini önemli ölçüde etkilemektedir [6]. Bir sınıflandırma

* Sorumlu yazar / Corresponding author, e-posta / e-mail: myilmaz@kayseri.edu.tr (M. Yılmaz)
Geliş / Recieved: 21.12.2023 Kabul / Accepted: 02.04.2024 Yayımlanma / Published: xx.xx.20xx
doi: 10.28948/ngumuh.1408200

yönteminin performansı, eğitim örneklerinin sayısı, özniteliklerin sayısı ve sınıflandırma yönteminin hesaplama karmaşıklığına bağlıdır [2]. Sınıflandırma yöntemlerinin hesaplama karmaşıklığı veri setinin boyutuna (n) göre genellikle üstel olarak artmaktadır. Bu sorunların çözümü için hiperspektral görüntülerde boyut indirgeme (Dimensionality Reduction) yöntemlerinin kullanılması gerekmektedir. Boyut indirgeme, hiperspektral görüntülerin veri analizinde önemli bir işlem adımıdır. Birçok araştırmacı, hiperspektral görüntüler için makul ve etkili bir boyut indirgeme yöntemi önerme girişimi üzerinde çalışmaktadır [7-10].

Boyut indirgemede özellik çıkarma ve özellik seçimi yöntemleri kullanılmaktadır. Özellik çıkarma ve özellik seçimi arasındaki seçim, uygulama alanına ve eğitim verilerine bağlıdır [7]. Özellik çıkarma, orijinal veri setindeki özelliklerin birleştirerek veya dönüştürerek yeni özelliklerin oluşturulmasıdır. Özellik seçimi ise bazı özellikleri doğrudan orijinal veri setinden seçer ve yeni bir özellik alt kümesi oluşturur [11].

Özellik çıkarımı, hiperspektral görüntünün kritik bilgilerini korur. Ancak, her bir spektral bandın fiziksel anlamını değiştirir [12]. Özellik çıkarımı, boyut indirgemenin gerekli olduğu durumlarda ve karmaşık sınıflandırma ve regresyon problemlerinde anlamlı sonuç üretme amacı ile kullanılabilir. Özellik seçimi ise yalnızca hiperspektral görüntünün kritik bilgilerini korumakla kalmaz, aynı zamanda her spektral bandın fiziksel anlamını da korur [12]. Bu nedenle hiperspektral görüntülerde fiziksel anlamının gerekli olduğu durumlar için daha uygundur. Çünkü orijinal özellikler değişmemiştir ve fiziksel anlamını korumaktadır [14].

Öznitelik seçme yöntemleri temel olarak seçilen bantların yeterli bilgi içermesi ve seçilen bantlar arasındaki korelasyonun düşük olması yaklaşımlarına odaklanır [15]. Özellik seçiminde kritik konu, olası bant kombinasyonlarından en iyi bantların seçimidir. Literatürde bant seçimi için özniteliklerin ağırlıkları büyükten küçüğe doğru sıralanarak ağırlığı büyük olan bandların seçimi benimsenmiştir. Ayrıca özellik alt kümelerinin hangisinin daha üstün olduğunu belirlemek için ise iteratif bir yaklaşımının daha iyi bir yol olduğu önerilmektedir [3].

Bant kombinasyonu seçimi belirsiz ve niceliksel bir problemdir [12]. Öznitelik seçimi öncesi ve sonrası oluşan hiperspektral görüntülerde arazi örtüsü sınıflarının spektral özellik dağılımında farklılıklar oluşur. Bu durum veri kümesi kayması (dataset shift) veya spektral kayma (spektrali shift) olarak adlandırılır [13]. Spektral kayma ile girdi ve çıktı verileri arasındaki temel ilişkiler değişebilir ve sınıflandırma modelinin performansı düşebilir.

Öznitelik seçme yöntemleri denetimli ve denetimsiz olmak üzere ikiye ayrılmaktadır. Denetimli öznitelik seçme yöntemleri genel olarak filtre, sarmalayıcı (Wrapper) ve gömülü (Embedded) yöntemler şeklinde sıralanmaktadır.

Filtre yöntemler, herhangi bir sınıflandırma algoritmasından bağımsızdır. Bu yöntemlerde özellikler, girdi ve hedef değişkeni arasındaki çeşitli istatistiksel ölçütlere göre seçilir. Bant değerlendirmesi için farklı ölçütler kullanılarak birçok filtre seçme yöntemi geliştirilmiştir. Amaç fonksiyonu mesafe, bilgi, korelasyon ve tutarlılık ölçütleri kullanılarak hesaplanır. Bilgi ölçüsü, filtreleme yöntemlerinde yaygın olarak kullanılan ölçütlerden biridir [14]. Filtreleme yöntemleri, modellerin eğitimini içermediğinden sarmalayıcı ve gömülü yöntemlere göre çok daha hızlıdır.

Sarmalayıcı yöntemler ise önceden belirlenmiş bir sınıflandırma algoritması gerektirir ve performansını değerlendirme kriteri olarak kullanır [1]. Bir özellik alt kümesi kullanarak eğitime başlar ve performansını değerlendirme kriteri kullanarak bir özellik ekler veya kaldırır. Durdurma kriterleri sağlanana kadar model eğitimini tekrarlar. Bu yöntemlerin yerel optimumlara takılma eğilimi vardır. Hiperspektral görüntülerde bantlar arasında güçlü bir korelasyon olduğu için iyi performans gösteremezler [1,16,17].

Sınıflandırmanın doğruluğunu ve etkinliğini artırmak için güvenilir, verimli ve etkili bir özellik seçim yöntemi kullanılmalıdır. Bu çalışmada sınıflandırma modelinden bağımsız olması, spektral bandın fiziksel anlamının korunması, hızı ve kararlığı nedeniyle Relief-F algoritması tercih edilmiştir.

Çalışmada, Relief-F yönteminin dört farklı ileri sınıflandırma algoritmasının performansına etkisi incelenmiştir. Bu algoritmaların performansı genel doğruluk, duyarlılık, kesinlik ve F1 puanı ile ispatlanmıştır. Ayrıca Relief-F algoritmasının performansı farklı veri boyutları ile test edilmiştir. Çalışma sonucunda Relief-F algoritmasının avantajları ve dezavantajları açıklanmıştır.

2 Materyal ve metot

2.1 Hiperspektral veri setleri

Bu çalışmada, veri setinin büyüklüğü ve özelliklerinin Relief-F algoritması üzerindeki etkisini incelemek için düşük, orta ve yüksek boyutlu örneklem büyüklüğüne sahip farklı sensörler tarafından elde edilen Salinas-A, Indian Pines ve Pavia University hiperspektral veri setleri kullanılmıştır.

Salinas, California Salinas vadisi üzerinde AVIRIS sensörleri tarafından toplanmış görüntülerden oluşmaktadır. 224 spektral bandı vardır. Bu spektral bantlardan 20 su emme bandı (108-112, 154-167, 224) çıkarılmıştır. 512×217 piksel, 54129 örnek ve 16 sınıftan oluşur. Salinas-A, Salinas veri setinin 204 spektral bant, 5348 örnek ve 6 sınıf ile 86 × 83 piksele sahip bir alt görüntüsüdür.

Indian Pines, Kuzeybatı Indiana'da AVIRIS sensörü tarafından toplanmış görüntülerden oluşmaktadır. 220 spektral bant, 10249 örnek ve 16 sınıf ile 145 × 145 piksele sahiptir. Bu spektral bantlardan 20 su emme bandı (104–108, 150–163 ve 220) çıkarılmıştır.

Pavia University, İtalya'nın kuzeyindeki Pavia üzerinde ROSIS sensörü tarafından toplanmış görüntülerden oluşmaktadır. 103 spektral bant, 42776 örnek ve 9 sınıf ile 610 × 340 piksele sahiptir. Hiperspektral veri setlerinin sınıf adları ve örnek sayıları Tablo 1'de verilmiştir.

Indian Pines veri seti dengesiz örneklem dağılımı nedeniyle çok zor bir sınıflandırma problemi olarak tanımlanabilir. Biçilmiş Çim-Mera ve Yulaf kategorileri için doğru şekilde sınıflandırılacak çok az sayıda örnek vardır.

Tablo 1. Hiperspektral veri setleri: sınıf adları ve örnek sayısı

SalinasA		Indian Pines		Pavia University	
Sınıf	Örnek	Sınıf	Örnek	Sınıf	Örnek
Brokoli (Yeşil Ot)	391	Yonca	46	Asfalt	6631
Mısır (Yeşil Ot)	1343	Mısır – işlenmemiş	1428	Çimen	18649
Romaine Cinsi Marul (4 hafta)	616	Mısır – az işlenmiş	830	Çakıl	2099
Romaine Cinsi Marul (5 hafta)	1525	Mısır	237	Ağaç	3064
Romaine Cinsi Marul (6 hafta)	674	Çim-Mera	483	Metal Levhalar	1345
Romaine Cinsi Marul (7 hafta)	799	Çim-Ağaçlar	730	Toprak	5029
		Biçilmiş, Çim ve Mera	28	Bitüm	1330
		Saman	478	Tuğla	3682
		Yulaf	20	Gölge	947
		Soya fasulyesi – işlenmemiş	972		
		Soya fasulyesi – az işlenmiş	2455		
		Soya fasulyesi	593		
		Buğday	205		
		Orman	1265		
		Binalar ,Çim, Ağaçlar, Yollar	386		
		Taş-Çelik-Kuleler	93		
Toplam	5348	Toplam	10249	Toplam	42776

Literatürde bazı araştırmacılar küçük örneklem büyüklüğüne sahip sınıfları veri setinden çıkarmaktadır [3].

2.2 Sınıflandırma yöntemleri

Sınıflandırma modelleri oluşturmak için yaygın olarak kullanılan ve en iyi sınıflandırma doğruluğu sağlayan Doğrusal Diskriminant Analizi (LDA), k-EYK, Rastgele

Orman (RO) ve DVM yöntemleri tercih edilmiştir. Dengesiz eğitim örneklerine karşı düşük hassasiyeti nedeniyle, DVM sınıflandırıcı hiperspektral görüntü analizinde kullanılan çok popüler bir yöntemdir [16].

2.2.1 Doğrusal disktriminant analizi

Diskriminant analizi, sınıflandırma problemleri için güçlü bir yöntemdir. Ancak, veri setinin doğru şekilde hazırlanması ve normal dağılım, çoklu bağlantı ve varyans kovaryans homojenliği gibi bazı istatiksel varsayımları sağlaması gerekmektedir. Bu varsayımlar sağlanması halinde doğrusal diskriminant fonksiyonu (LDA) kullanılır. Verilerin normal dağılımlı olduğu ve varyans kovaryans eşitliği olmadığı durumlarda ise ikinci dereceden diskriminant fonksiyonu (QDA) kullanılır [18].

2.2.2 Rastgele orman

Rastgele Orman (RO) algoritması, veri setinden rastgele örnekleme yapar ve her bir ağaç için farklı bir veri alt kümesi oluşturur. Bu işlem aşırı öğrenmeyi engellemek için kullanılır. Her alt küme üzerinde birçok karar ağacı oluşturulur ve karar ağaçları tahminler yapar. Daha sonra tüm karar ağaçlarının tahminleri birleştirilerek daha doğru sonuçlar elde edilemeye çalışılır [19].

2.2.3 K-en yakın komşuluk

k-EYK, sınıflandırma problemlerini çözmek için basit ve etkili bir algoritmadır. k-EYK algoritması, veri setini sınıflandırmak için bir uzaklık kriteri kullanır. Bu kriter, bir girdi değeri ile eğitim veri setindeki diğer veriler arasındaki benzerliği veya uzaklığı ölçer. Ayrıca algoritma, sınıflandırmada kullanılacak komşu sayısını belirlemek için bir K değeri belirlenmelidir. K değeri, problemin özelliğine göre seçilir [20].

2.2.4 Destek vektör makinesi

Destek vektör makinesi (DVM), bir hiper düzlem kullanarak sınıflandırma işlemini gerçekleştirir. Yöntem veri setini iki veya daha fazla sınıfı en iyi şekilde bölen bir hiper düzlemi bulmayı amaçlar. Günlük hayatta kullanılan veriler genellikle doğrusal olmadığı için bu tip verilerde DVM yönteminin kullanılabilmesi için kernel (çekirdek) fonksiyonlarına ihtiyaç duyulmaktadır. Modelin eğitimi sırasında aşırı öğrenmeyi engellemek için kutu kısıtlaması ile destek vektörlerinin sayısı ayarlanabilmektedir. Ayrıca girdi verilerinin kernel işlevine uygulanmadan önce bir özelliğe göre ölçeklendirilmesi önerilir. Bazı özelliklerin mutlak değerleri geniş bir aralıkta olduğunda veya büyük olabildiğinde, bunların iç çarpımı çekirdek hesaplamasında baskın olabilir. Bunu önlemek için kernel ölçeği kullanılmaktadır [21].

2.3 Relief-F algoritması

Relief-F filtre tabanlı denetimli öznitelik seçim algoritmasıdır. Relief-F algoritması özelliklerin kalitesini, değerlerinin birbirine yakın örnekler arasında ne kadar iyi ayrım yaptığına bağlı olarak en iyi nitelikleri tahmin eder [22]. Bu nedenle Relief-F büyük ölçüde veri örneklerine bağlıdır [12]. Relief-F algoritmasının işlem adımları Tablo 2'de verilmiştir. Algoritmada rastgele bir R_i örneği seçilir. Daha sonra aynı H sınıfından k en yakın komşuyu (nearest hits) ve farklı M sınıflarının her birinden k en yakın komşuyu (nearest misses) belirler. En yakın komşu sayısı (k), kullanıcı tarafından tanımlanır. k parametresi tahminlerin yerini kontrol eder ve gürültü ile ilgili algoritmanın daha sağlam olmasını sağlar. R_i,M ve H değerlerine bağlı olarak tüm F özelliklerin ağırlıkları W[A] güncellenir. Eğer bir F özelliği, aynı sınıfa ait olmasına rağmen, R'yi H'den ayırmak için kullanılabiliyorsa ağırlık tahmini W(F) azaltılır. Aksine, R'yi

Tablo 2. Relief-F algoritmasının işlem adımları

Relief-F Algoritması	
Girdi:	Her eğitim örneği için özellik değerlerinin ve sınıf değerinin bir vektörü
Output:	Öznitelikler için tahmin edilen W (ağırlık) vektörü
1	Başlangıç ağırlıkları W[A]=0;
2	for i=1:m
3	Rastgele bir Ri örneği seç
4	Aynı sınıftan k tane en yakın örneği Hj bul
5	for C≠sınıf(Ri)
6	Aynı sınıfta olmayan diğer bütün sınıflardan k tane en yakın örneği Mj(C) bul
7	for A=1:a
8	$W[A] = W[A] - \sum_{j=1}^{k} \frac{diff(A,R_i,H_j)}{(m.k)} + \sum_{C \neq sınıf(R_i)} \frac{\left[\frac{P(C)}{1-P(sınıf(R_i))} \sum_{j=1}^{k} diff\left(A,R_i,M_j(C)\right)\right]}{(m.k)}$
9	end

M'den ayırmak için bir F özelliği kullanılabiliyorsa, ağırlık tahmini W(F) artırılır. Yani algoritma F özelliklerini benzer sınıfları ayırmadığı için ödüllendirirken, F özelliklerini farklı sınıfları ayırmadığı için cezalandırmaktadır. Tüm aynı sınıftaki k en yakın komşu ve farklı sınıflardaki k en yakın komşuların katkısının ortalaması alınır ve o sınıfın eğitim veri setinden tahmin edilen önceki olasılığı ile ağırlıklandırılır. Bu işlem m kez tekrarlanır. Burada m, iterasyon sayısı olarak kullanıcı tanımlı bir parametredir. Çıktılar ise her bir özelliğin önemini gösteren ağırlıklardır [23].

Robnik-Sikonja ve Kononenko [22] çalışmalarında komşu sayısının genel olarak 10'a ayarlanabileceğini belirtmişlerdir. Ancak problemin özelliğine, problemin karmaşıklığına, gürültü miktarına ve örnek sayısına göre değişebilmektedir. Komşu sayısının değeri çok yükselirse, kalite tahmini W(F)'nin pozitif ve negatif güncellemeleri eşitlenebilir hale gelir ve W(F) sıfıra ulaşır. Önemli özellikler, önemsiz özelliklerden ayırt edilemez hale gelir.

Relief-F algoritmasında tüm k en yakın örneğinin sabit bir etkisi kullanılmaktadır. Buna göre mesafe etkisi ağırlık güncelleme denkleminde d(i, j) = 1/k şeklinde tanımlıdır. Denetimli öğrenmede mesafeye göre ağırlıklandırmanın yararlı olduğu düşünülmektedir. Relief-F algoritmasında k en yakın komşunun mesafesini (d) dikkate almak için ağırlık güncelleme denklemi Denklem (1)'deki gibi değiştirilmektedir [22].

$$W[A] = W[A] - \frac{1}{m}\sum_{j=1}^{k} diff(A,R_i,H_j)d(R,H_j) + \frac{1}{m}\sum_{C \neq sınıf(R_i)} \frac{P(C)}{1-P(sınıf(R_i))}\sum_{j=1}^{k} diff\left(A,R_i,M_j(C)\right)d(R,M_j(C)) \quad (1)$$

Algoritma gerçek mesafeleri kullanılırken en yakın komşu sayısına karşı duyarlıdır. Bu nedenle sonuçlarda önemli sapmalar olabilir. Gerçek mesafeler kullanılırken en yakın komşu sayısı için farklı değerler test edilmelidir. İterasyon sayısı 20 ile 50 arasında ayarlanmalıdır. Tahminleri hassaslaştırmak için daha fazla iterasyon sayısı ayarlanabilir. Ayrıca iterasyon sayısı probleme bağlıdır [22].

Özniteliklerin kalitesini tahmin etmek için kullanılan ölçütlerin çoğu özelliklerin ve hedef değişkeni üzerinde bağımsızlığını varsayarlar. Relief-F algoritması ise bu varsayımı yapmaz. Bu nedenle nitelikler arasında güçlü korelasyonun olduğu problemlerde niteliklerin kalitesini doğru bir şekilde tahmin edebilir [1,22]. Ayrıca gürültülü ve eksik verileri de işleyebilir [22].

Relief-F algoritmasının yukarıdaki avantajları dışında bazı dezavantajları da vardır. Bunları şu şekilde sıralayabiliriz.

- Denetimli bir algoritma olması nedeniyle az sayıda örnek içeren yüksek boyutlu veri setlerinde optimum şekilde performans göstermeyebilir [22].
- Özellik sayısı çok büyük olduğunda hesaplama açısından zaman alıcıdır.
- Gereksiz özellikler mevcut olduğunda tahminlerinin performansı düşebilir.

Öznitelik alt kümesi seçimi, hedef değeri açıklamak için gerekli ve yeterli en küçük bir küme seçme problemidir. Kira ve Rendell [23], en önemli özellikleri seçmek için eşik değer olarak Denklem (2)'deki h değerini önermektedir.

$$0 < h < \frac{1}{\sqrt{\alpha m}} \quad (2)$$

Denklemde α önemsiz bir özelliği kabul etme olasılığı ve m kullanılan iterasyon sayısıdır.

2.4 Performans metrikleri

Relief-F algoritmasının performansını değerlendirmek için hata matrisinden (confusion matrix) hesaplanan genel sınıflandırma doğruluğu, duyarlılık, kesinlik, F1 puanı ve kappa katsayısı metrikleri kullanılmıştır. Bu metriklerin formülleri ve açıklamaları Tablo 3'te sunulmuştur.

Dengesiz örneklem dağılımı olan veri setlerinde model doğruluğu tek başına yeterli değildir. Bu nedenle doğruluk metriği F1 puanı [24] ve kappa katsayısı [25] metrikleri ile birlikte değerlendirilmedir. F1 puanı sadece modelin tahmin hatalarının sayısını değil, aynı zamanda yapılan hataların türünü de dikkate alır. Hem kesinlik hem de duyarlılık dengelemek istenildiğinde bu metrik kullanılır [24].

Tablo 3. Performans metriklerinin formül ve açıklamaları

Performans Metriği	Formül	Açıklama
Genel Doğruluk	(TP+TN)/(TP+FP+TN+FN)	Doğru tahminlerin toplam tahminlere oranını
Duyarlılık	TP/(TP+FN)	Doğru tespit edilen pozitif sınıfların oranı
Kesinlik	TN/(TN+FP)	Doğru tespit edilen negatif sınıfların oranı
F1 puanı	2*[(Duyarlılık* Kesinlik) / (Duyarlılık+Kesinlik)]	Kesinlik ve duyarlılık değerlerinin harmonik ortalaması
Kappa katsayısı	(Gözlemlenen doğruluk - Beklenen doğruluk)/(1 - Beklenen doğruluk)	Gözlenen doğruluğu, beklenen doğrulukla (rastgele şans) karşılaştıran bir ölçüdür.

TP: Gerçek Pozitif, FP: Yanlış Pozitif, TN: Gerçek Negatif, FN: Yanlış Negatif, Gözlemlenen doğruluk: Doğruluk,
Beklenen doğruluk: [(TN+FP)*(TN+FN)+(FN+TP)*(FP+TP)]/[N*N]

Kappa istatistiği, bir hata matrisinin doğru yüzde değerlerinin (TP+TN) yer gerçeği (ground truth) verileri ile ne kadar yakından eşleştiğinin bir ölçüsüdür. Kappa istatistiği, az sayıda örneklem olan sınıfları daha çok dikkate alması nedeniyle model performansına ilişkin daha gerçekçi bir değer üretir [25].

Salinas-A, Indian Pines ve PaviaU veri setlerinden Relief-F algoritması ile 5 band aralığında band altküme seçimi yapılmış ve LDA, k-EYK, RO ve DVM sınıflandırma yöntemlerin performansları karşılaştırılmıştır. Üç veri setinin girdi değişkenleri, özellik seçimi uygulanmadan önce sıfır ortalama değere ve birim varyansa sahip olacak şekilde standartlaştırılmıştır.

Veri setlerinin rastgele etkisini azaltmak için hiperspektral görüntüler 3 kat çapraz doğrulama ile eğitim ve test veri seti olarak ikiye ayrılmıştır. Her sınıf için düşük örneklem büyüklüğü nedeniyle 3 kat çapraz doğrulama kullanılmıştır. Farklı sınıflandırma yöntemleri arasında anlamlı bir karşılaştırma yapmak için, aynı eğitim ve test verileri ile modeller oluşturulmuştur. Sınıflandırma yöntemlerinin hiper parametreleri yazarların tecrübelerine göre belirlenmiştir [26]. Sınıflandırma yöntemlerinde kullanılan parametreler Tablo 4'de verilmiştir.

Tablo 4. Sınıflandırma yöntemlerinin parametreleri

Sınıflandırma yöntemi	Parametreler
LDA	Diskriminant fonksiyonu: Quadratik Gama Katsayısı: 0
k-EYK	Uzaklık kriteri: Öklid, Komşu sayısı: 1
RO	Bölme kriteri: Gini indeksi Maksimum karar bölme sayısı: n*-1
DVM	Çekirdek fonksiyonu: Gaussian, Kernel Ölçeği: 5, Kutu Kısıtlaması: 5

n: Eğitim seti büyüklüğü

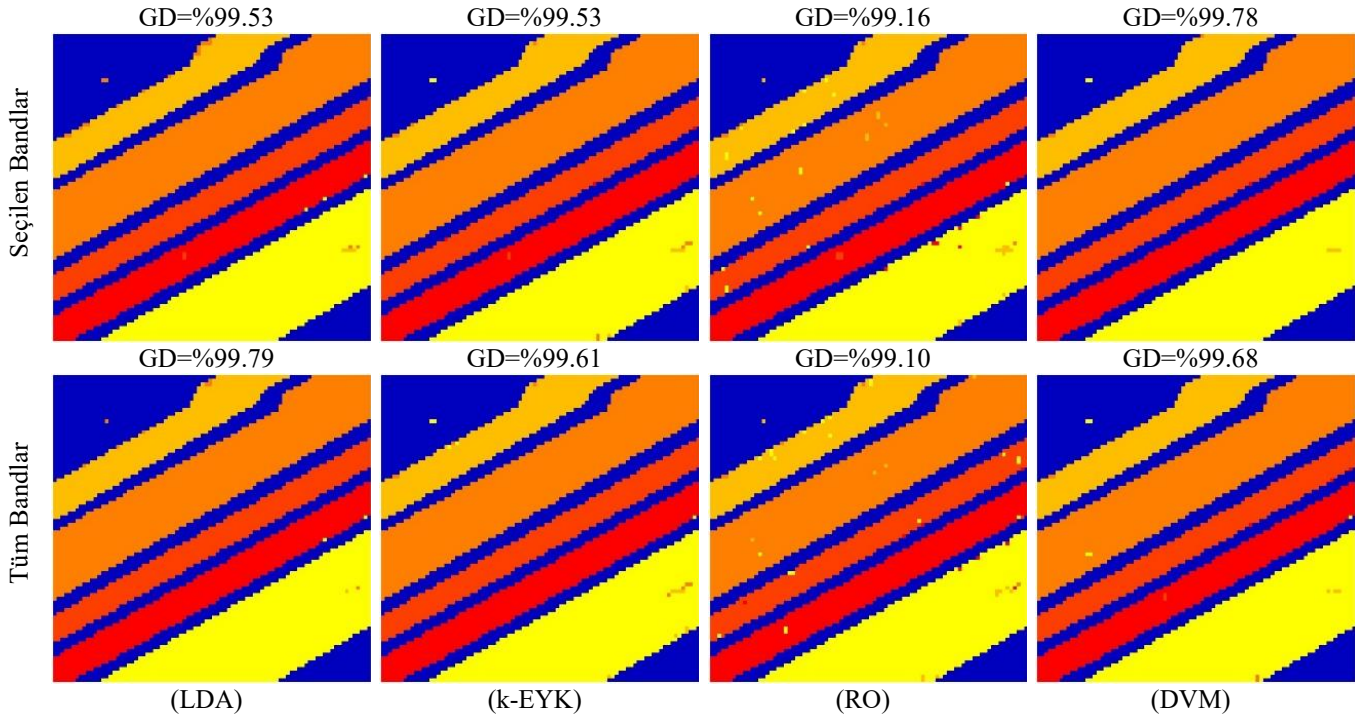
Relief-F algoritmasında, bantlar hesaplanan önem puanlarına göre sıralanır ve veri boyutunu azaltmak için en üstteki k adet bant seçilir. Relief-F algoritmasının iterasyon sayısı (m) ve komşu sayısı (k) olmak üzere belirlenmesi gereken iki parametresi vardır. İterasyon sayısı, eğitim veri setindeki örneklerin sayısı olarak ayarlanmıştır. Yani bandların ağırlıklarını belirlemek için tüm örnekler kullanılmıştır. Komşu sayısı parametresi için literatürde yaygın olarak kullanılan 10 değeri ayarlanmıştır. Relief-F algoritmasında mesafe fonksiyonu olarak Manhattan mesafesi seçilmiştir. Bu parametreler ile bantların ağırlık matrisi elde edilmiştir ve veri setlerinin tüm bandları ağırlıklarına göre azalan şekilde sıralanmıştır. SalinasA ve Indian Pines veri seti için ilk 50 band, PaviaU veri seti için ise ilk 40 band seçildi. Seçilen band sayılarına göre sınıflandırma sonuçları Şekil 1, 2, 3'te verilmiştir.

Şekil 4, çalışmamızda kullanılan üç veri seti ve tüm yöntemler için band sayısı ile genel sınıflandırma doğruluğunun değişimini göstermektedir. Tüm veri setlerinde en iyi doğruluk değerleri DVM yöntemi ile elde edilmiştir. Grafikte görüldüğü gibi tüm durumlarda seçilen band sayısı artıkça genel sınıflandırma doğruluğunu da artmaktadır. Ancak SalinasA ve Indian Pines veri setlerinde ilk 50 bandın kullanıldığı sonuçlar ile tüm bandların kullanıldığı sonuçlar arasında istatiksel olarak anlamlı bir fark bulunmamaktadır. PaviaU verisetinde ise ilk 40 banttan sonra genel doğruluk (GD) değerlerinde anlamlı bir gelişme olmamaktadır.

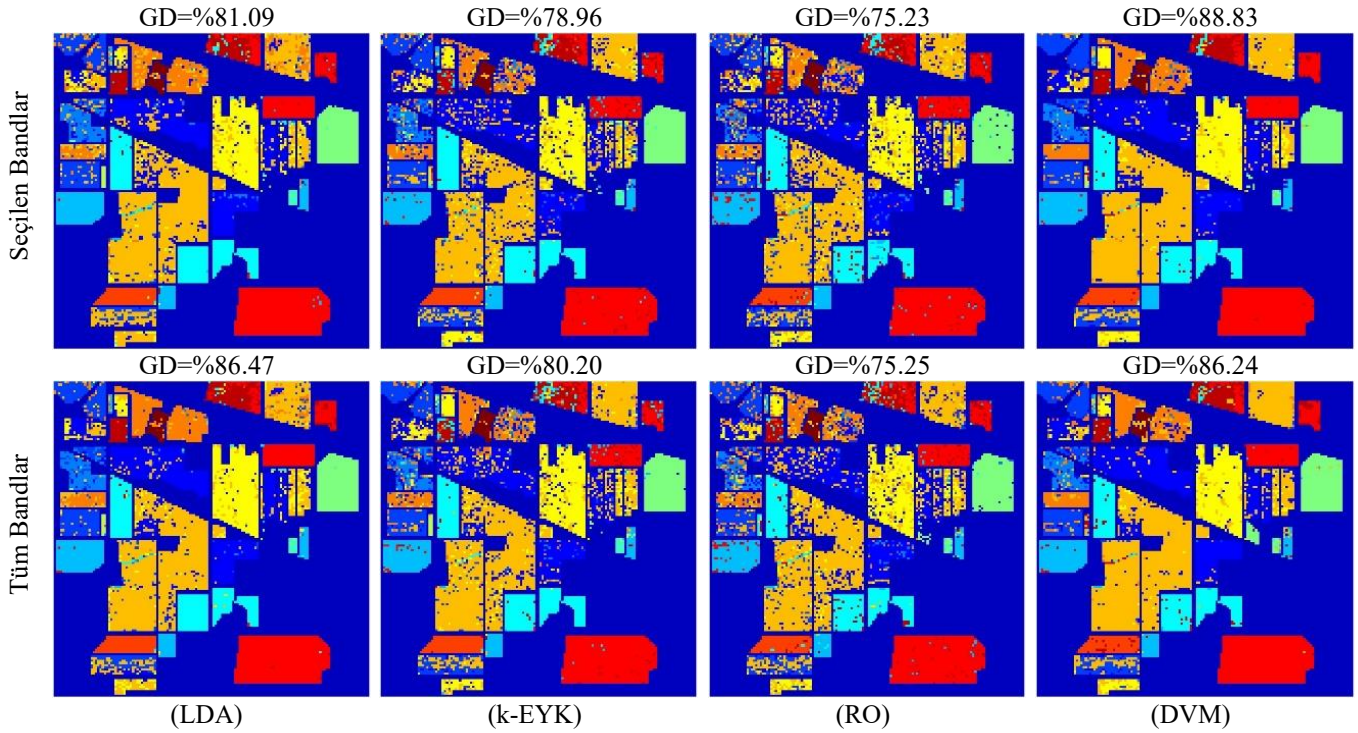
SalinasA, Indian Pines ve PaviaU veri setleri için seçilen bant sayısı sırasıyla 50, 50 ve 40'tır. Seçilen band sayıları ile orijinal veri setlerine göre veri boyutunu azaltılmıştır. Veri boyutunun azalmasıyla orantılı olarak sınıflandırma yöntemlerinin hesaplama karmaşıklığı da düşmektedir. Hiperspektral görüntülerde sınıflandırma verimliliğini artırmak için iyi bir sınıflandırma doğruluğu ile bant sayısının %60 - %70 azaltılabileceği görülmektedir.

Sınıflandırma sonuçları genel doğruluk, duyarlılık (sensitivity), kesinlik (precision), F1 puanı ve kappa katsayısı ölçütlerine göre değerlendirilmektedir. Üç veri setindeki DVM sınıflandırma yönteminin performans metrikleri Şekil 5'te verilmiştir.

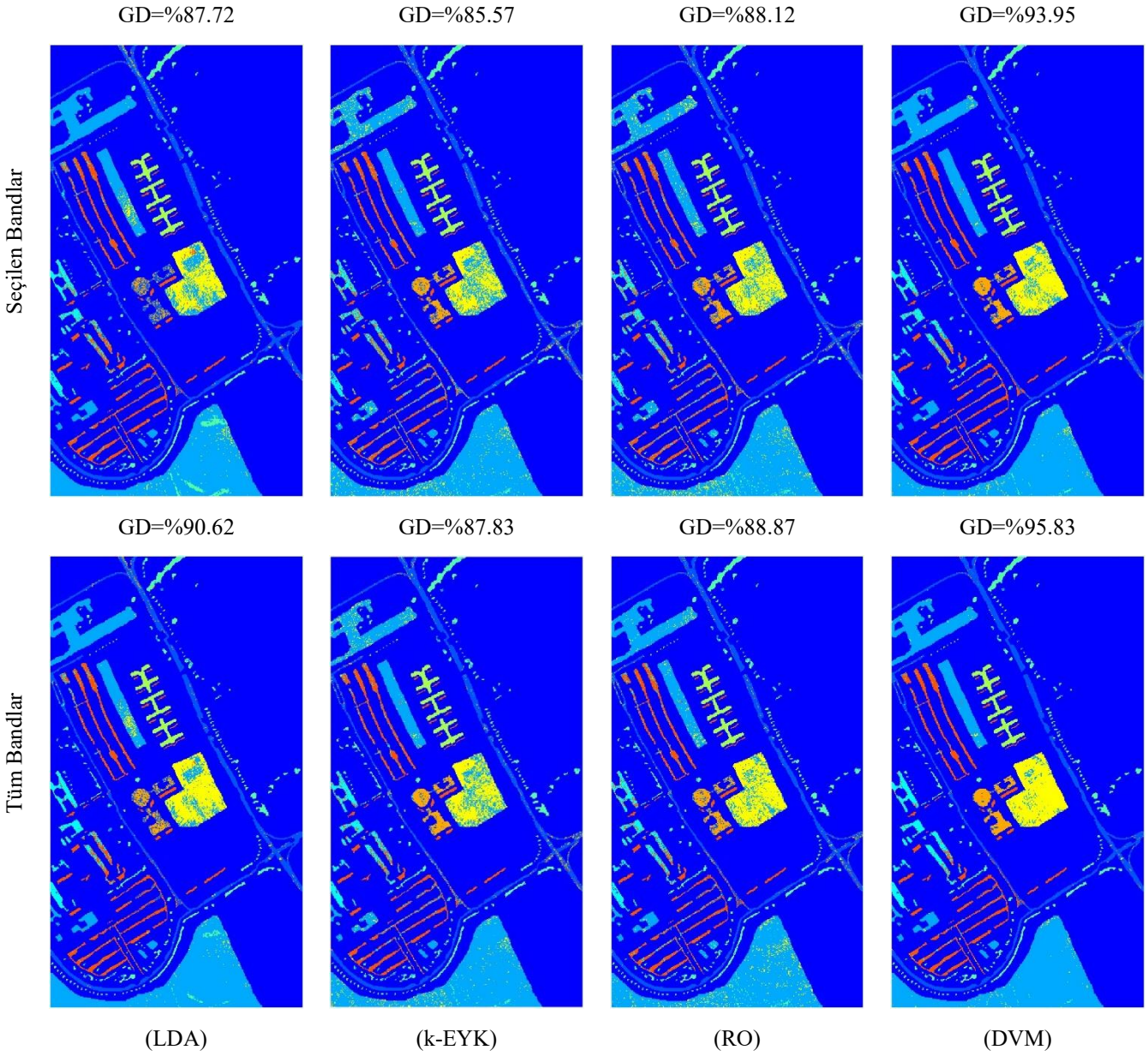
Kappa katsayısı, doğru sınıflandırılmış piksellerin değerleri ile yer gerçeği piksellerinin değerleri arasındaki uyumu hesaplar. Kappa katsayısında genel doğruluk değerinden farklı olarak şans faktörünün etkisi azaltılmaya çalışılmaktadır. Bu nedenle Şekil 5'te görüldüğü gibi kappa katsayıları genel doğruluk değerine göre daha düşük değerlerdir. SVM yönteminin genel doğruluğu ağırlık değerlerine göre ilk 50 banttan sonra oldukça yüksektir. Indian pines veri seti bazı sınıflarda çok sınırlı sayıda örneklem içerdiği için diğer veri setlerine göre daha az genel doğruluk değeri ile band seçimi yapılabilmiştir.



Şekil. 1. Seçilen ve tüm bandlar ile elde edilen sınıflandırma görüntüleri (SalinasA)



Şekil. 2. Seçilen ve tüm bandlar ile elde edilen sınıflandırma görüntüleri (Indian Pines)



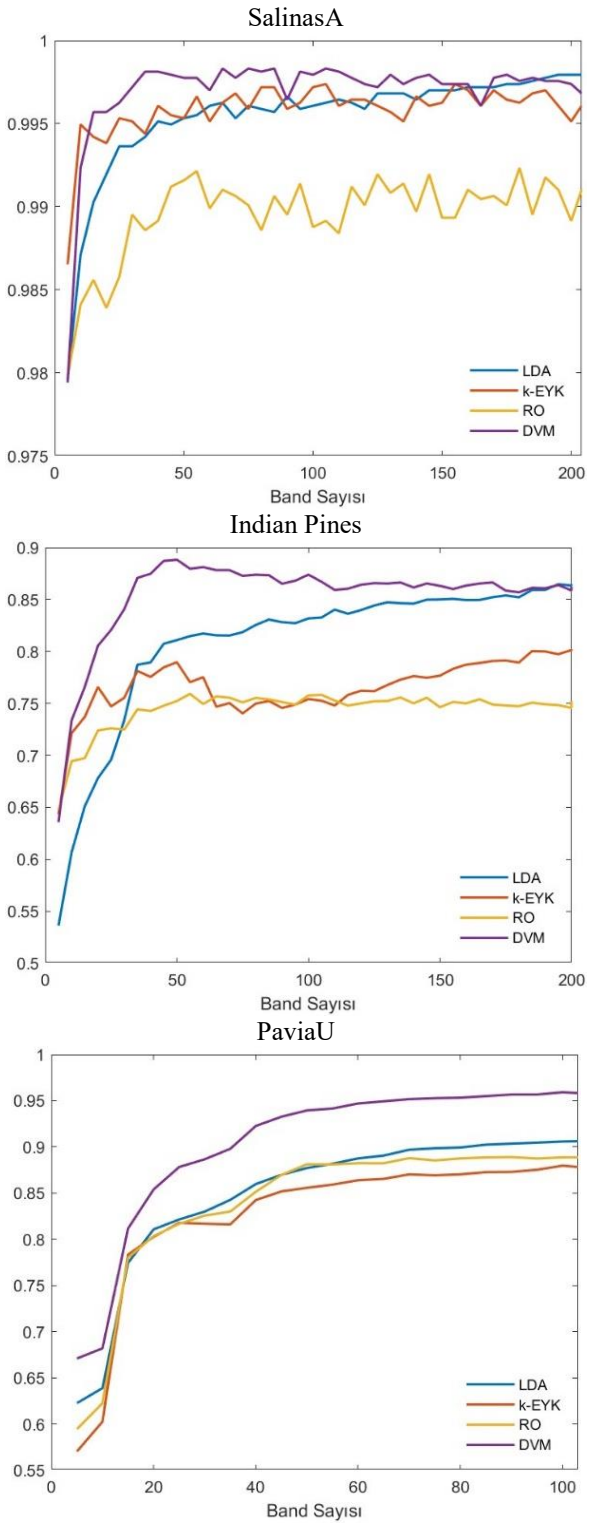
Şekil. 3. Seçilen ve tüm bandlar ile elde edilen sınıflandırma görüntüleri (PaviaU)

SalinasA, Indian Pines ve PaviaU veri setlerinin kappa değerleri sırasıyla 0.9972, 0.8721 ve 0.9194 şeklindedir. Indian Pines veri setinde kappa değerinin 0.87 olması, gözlemlenen bir sınıflandırmanın şansa dayalı bir sınıflandırmadan %87 daha iyi olduğunu göstermektedir.

Seçilen bandlar, tüm bandlar ile elde edilen sınıflandırma doğruluğuna oldukça yakın değerler sağlamaktadır. Küçük boyutlu veride (SalinasA) band sayısı artıkça doğrulukta iyileşme olmamaktadır. Büyük boyutlu veride (PaviaU) band sayısı arttıkça doğrulukta küçük bir iyileşmenin olduğu görülmektedir. Ancak tüm bandlar kullanıldığında sınıflandırma süresi oldukça artmaktadır. Dengesiz dağılımlı veride ise band sayısı arttıkça ilk 50 banttan sonra sınıflandırma doğruluğu düşmektedir.

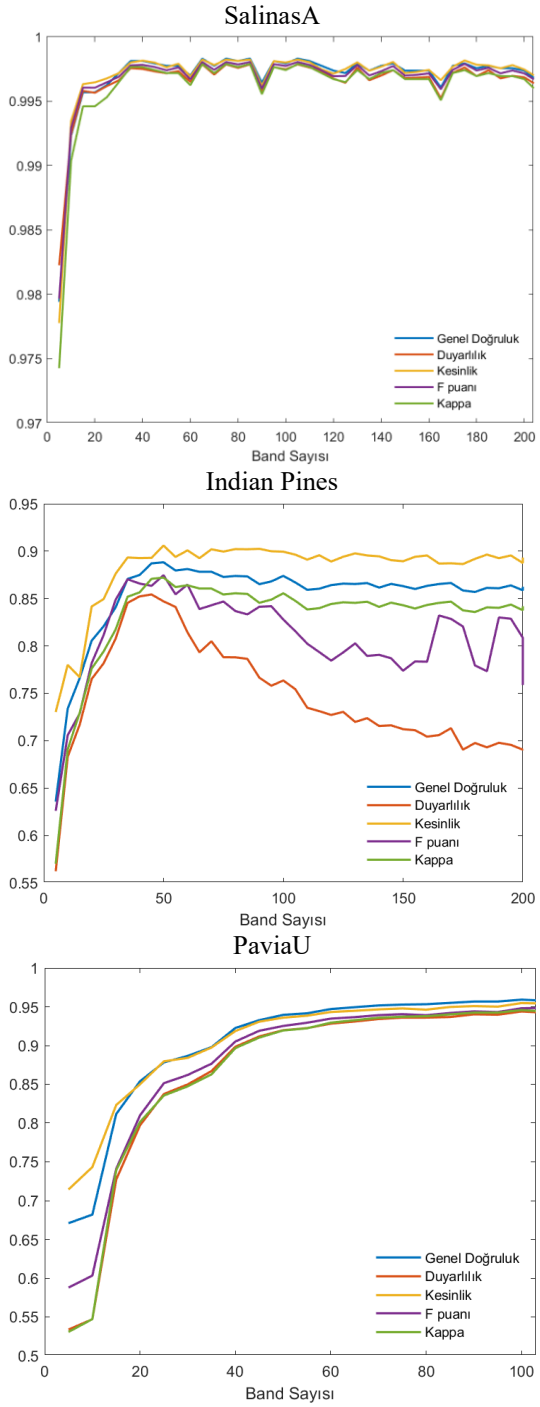
Relief-F algoritmasının büyük boyutlu verilerde sınıflandırma süresini iyileştirdiği, dengesiz dağılımlı verilerde ise sınıflandırma doğruluğunu iyileştirdiği söylenebilir. Şekil 1'de verildiği gibi SalinasA veriseti ile yapılan uygulamada DVM yönteminin seçilen bandlar ile gerçekleştirilen uygulamada, tüm bandların kullanıldığı uygulamaya göre küçük bir farkla daha başarılı olduğu; diğer yöntemlerde ise seçilen bandlar ile tüm bandların kullanıldığı uygulamalarla çok yakın sonuçlar elde edildiği görülmüştür. Indian Pines veriseti ile gerçekleştirilen uygulamada DVM yönteminin band seçimi sonrası doğruluğunun arttığı, RO yönteminin ise doğruluğunun çok büyük oranda korunduğu görülmüştür. PaviaU veriseti ile yapılan uygulamada ise RO yönteminin büyük oranda doğruluğunun korunduğu; diğer yöntemlerde ise yaklaşık

%1.5-2 arasında doğruluğun düştüğü anlaşılmıştır. DVM yönteminin Relief-F yöntemi ile iki verisetinde (SalinasA ve Indian Pines) başarı artışı sağlanmıştır. PaviaU verisetinin sınıflarının dengesiz bir dağılıma sahip olması, band seçim işleminin sınıflandırma performansına etkisinin istenen seviyelere ulaşmasını zorlaştırmıştır.

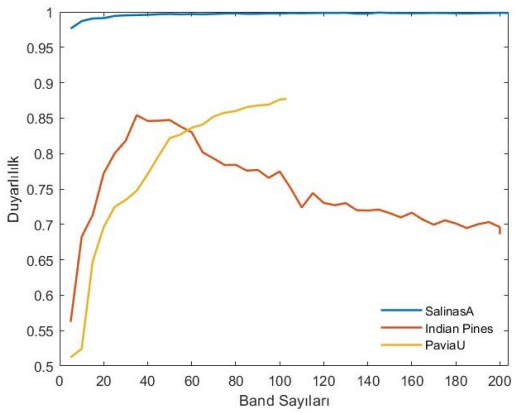


Şekil. 4. Veri setlerinin band sayılarına ve sınıflandırma yöntemlerine göre genel doğruluk değerleri

Üç veri setindeki DVM sınıflandırma yönteminin duyarlılık metrikleri Şekil 6'da verilmiştir. SalinasA veri setinde ilk 50 band ile tüm bandların duyarlılığı yaklaşık aynıdır. Indian Pines veri setinde ilk 50 banttan sonra duyarlılık düşmektedir. PaviaU veri setinde ise band sayısı arttıkça duyarlılık değeri artmaktadır. Buna göre Relief-F algoritması ile çalışmada kullanılan SalinasA, Indian Pines ve PaviaU veri setlerinde ilk 50 bantta orijinal verinin çoğu özelliğinin korunduğu görülmektedir.

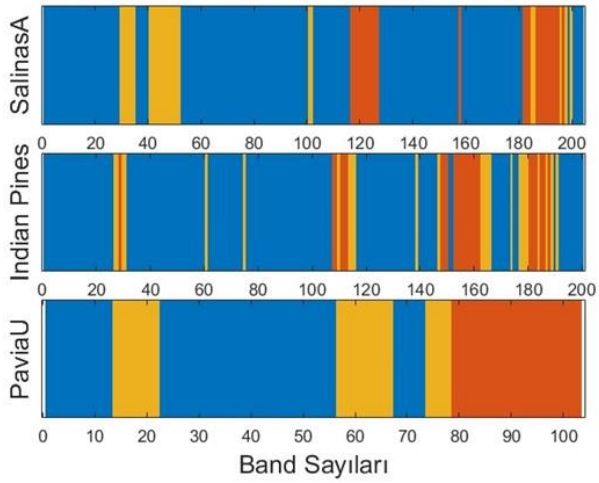


Şekil. 5. Veri setlerinin DVM yöntemindeki performans metrikleri



Şekil 6. Veri setlerinin DVM yöntemine göre duyarlılık değerleri

Relief-F tarafından seçilen özelliklerin genel olarak bitişik bantlar olduğu Şekil 7'de görülebilmektedir. Komşu bantlar arasındaki korelasyon diğer bantlara göre daha yüksektir. Algoritma tarafından benzer korelasyonlu bantlara benzer ağırlıklar hesaplanmıştır. Şekildeki turuncu kısım ağırlık değerlerine göre ilk 25 bandı, sarı kısım 26'dan 50'ye kadar puan alan bantları ve mavi kısım ise diğer bantları temsil etmektedir. Ağırlıklarına göre en önemli bantlar, yakın kızılötesi (NIR; 0.75-1.4 µm) ve kısa dalga kızıl ötesi (SWIR; 1.1 – 3.0) bölgelerindeki bantlardır.



Şekil 7. Relief-F ile seçilen bandların haritası

4 Sonuçlar ve gelecek çalışmalar

Hiperspektral görüntüler, detaylı spektral bilgiye sahiptir. Ancak verilerin band sayısı ve örneklem büyüklüğü arttıkça sınıflandırma sonuçları olumsuz etkilenmektedir. Öznitelik seçim yöntemleri ile hiperspektral görüntülerin sınıflandırma doğruluğu iyileştirilebilir. Bu çalışmada, hiperspektral görüntülerin band seçiminde Relief-F algoritmasının etkinliği, sınıf dağılımının hem düzgün olduğu hem de dengesiz olduğu deneysel veri setleri üzerinde ayrı ayrı test edilmiştir ve sonuçlar güncel ileri metriklerde ortaya konulmuştur. Relief-F algoritması, araştırmamızda kullanılan üç hiperspektral veri setinin tümü için genel sınıflandırma doğruluğu ve Kappa katsayısı açısından önemli bir gelişme göstermiştir. Deneysel sonuçlar, Relief-F algoritmasının, bantlar arasındaki fazlalığı etkili bir şekilde azaltabileceğini göstermektedir.

Relief-F sınıflandırma yönteminden bağımsız olarak hedef değerlere göre öznitelik ağırlıklarını belirlemektedir ve bu ağırlıklara göre band seçimi yapılmaktadır. Seçilen bandlar ile tüm bandların sınıflandırma sonuçları karşılaştırıldığında, Relief-F algoritması sınıflandırma doğruluğuna ciddi katkısı olmayan bandları kaldırabilir ve hesaplama yükünü azaltabilir. Ayrıca Relief-F, hem hassasiyet hem de özgüllük (specificity) metriklerini dikkate alarak öznitelik seçimini yapmaktadır. Bu durum, hiperspektral görüntünün fiziksel özelliğinin korunmasına yardımcı olabilir.

Relief-F algoritmasının dezavantajları olarak öznitelik seçim stratejisindeki bazı eksiklikleri tespit edilmiştir. Bu eksiklikleri gidermek için gelecek çalışmalarda aşağıdaki işlemler yapılabilir. Relief-F algoritması ile seçilen bandlar arasında güçlü korelasyon devam etmektedir. Bandlar arasındaki korelasyonu dikkate alan diğer özellik seçimi yöntemleri ile birleştirilerek hibrit özellik seçimi yapılabilir. Relief-F algoritması, çok büyük veri setlerinde hesaplama açısından işlem yükü fazladır. Ağırlık güncellemesinde bütün örneklerin kullanılması yerine bir optimizasyon algoritması geliştirilerek işlem süresi azaltılabilir. Ayrıca gelecek çalışmalarda farklı mesafe ölçütlerinin Relief-F algoritmasının performansına etkisi incelebilir ve yakın ve uzak örneklerin ayırt edilebilirliğini artırmak için çekirdek (kernel) tabanlı bir Relief-F algoritması tasarlanabilir.

Çıkar çatışması

Yazarlar çıkar çatışması olmadığını beyan etmektedir.

Benzerlik oranı (iThenticate): %4

Kaynaklar

[1] A. Ghosh, A. Datta and S. Ghosh, Self-adaptive differential evolution for feature selection in hyperspectral image data. Applied Soft Computing, 13 (4), 1969-1977, 2013. https://doi.org/10.1016/j.asoc.2012.11.042.

[2] S. Y. Xiang, Z. H. Xu, Y. W. Zhang, Q. Zhang, X. Zhou, H. Yu, B. Li and Y. F. Li, Construction and Application of ReliefF-RFE Feature Selection Algorithm for Hyperspectral Image Classification. Spectroscopy and Spectral Analysis, 42 (10), 3283-3290, 2022.

[3] B. Wu, C. C. Chen, T. M. Kechadi and L. Y. Sun, A comparative evaluation of filter-based feature selection methods for hyper-spectral band selection. International Journal of Remote Sensing, 34 (22), 7974-7990, 2013. https://doi.org/10.1080/01431161.2013.827815.

[4] J. S. Ren, R. X. Wang, G. Liu, R. Y. Feng, Y. N. Wang and W. Wu, Partitioned Relief-F Method for Dimensionality Reduction of Hyperspectral Images. Remote Sensing, 12 (7), 21, 2020. https://doi.org/10.3390/rs12071104.

[5] T. Lillesand, R. W. Kiefer and J. Chipman, Remote Sensing and Image Interpretation. Wiley, 2015.

[6] B. Rasti, D. F. Hong, R. L. Hang, P. Ghamisi, X. D. Kang, J. Chanussot and J. A. Benediktsson, Feature Extraction for Hyperspectral Imagery: The Evolution From Shallow to Deep: Overview and Toolbox. Ieee Geoscience and Remote Sensing Magazine, 8 (4), 60-88, 2020. https://doi.org/10.1109/mgrs.2020.2979764.

[7] R. Jung and M. Ehlers, Comparison of two feature selection methods for the separability analysis of intertidal sediments with spectrometric datasets in the German Wadden Sea. International Journal of Applied Earth Observation and Geoinformation, 52, 175-191, 2016. https://doi.org/10.1016/j.jag.2016.06.009.

[8] Y. Dong, B. Du, L. Zhang and L. Zhang, Dimensionality Reduction and Classification of Hyperspectral Images Using Ensemble Discriminative Local Metric Learning. IEEE Transactions on Geoscience and Remote Sensing, 55 (5), 2509-2524, 2017. https://doi.org/10.1109/TGRS.2016.2645703.

[9] X. Zhang, X. Jiang, J. Jiang, Y. Zhang, X. Liu and Z. Cai, Spectral–Spatial and Superpixelwise PCA for Unsupervised Feature Extraction of Hyperspectral Imagery. IEEE Transactions on Geoscience and Remote Sensing, 60, 1-10, 2022. https://doi.org/10.1109/TGRS.2021.3057701.

[10] M. R. Islam, A. Siddiqa, M. Ibn Afjal, M. P. Uddin and A. Ulhaq, Hyperspectral Image Classification via Information Theoretic Dimension Reduction. 15 (4), 1147, 2023.

[11] X. C. Y. Su and F. Liu, A Survey For Study of Feature Selection Based On Mutual Information. 9th Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (WHISPERS), pp. 1-4, Amsterdam, Netherlands, 2018.

[12] Z. H. Wang, S. L. Liang, L. Z. Xu, W. Song, D. X. Wang and D. M. Huang, Dimensionality reduction method for hyperspectral image analysis based on rough set theory. European Journal of Remote Sensing, 53 (1), 192-200, 2020. https://doi.org/10.1080/22797254.2020.1785949.

[13] M. C. Ye, Y. Q. Xu, C. X. Ji, H. Chen, H. J. Lu and Y. T. Qian, Feature selection for cross-scene hyperspectral image classification using cross-domain ReliefF. International Journal of Wavelets Multiresolution and Information Processing, 17 (5), 17, 2019. https://doi.org/10.1142/s0219691319500395.

[14] A. Elmaizi, E. Sarhrouni, A. Hammouch and C. Nacir, A new band selection approach based on information theory and support vector machine for hyperspectral images reduction and classification. International Symposium on Networks, Computers and Communications (ISNCC), pp. 1-6, Marrakech, Morocco, 2017. https://doi.org/10.1109/ISNCC.2017.8072002.

[15] S. Zhou, J. P. Zhang and B. K. Su, Feature selection and classification based on ant colony algorithm for hyperspectral remote sensing images. 2nd International Congress on Image and Signal Processing, pp. 1-4, Tianjin, China, 2009. https://doi.org/10.1109/CISP.2009.5304614.

[16] W. W. Sun and Q. Du, Hyperspectral Band Selection A review. Geoscience and Remote Sensing Magazine, 7 (2), 118-139, 2019. https://doi.org/10.1109/mgrs.2019.2911100.

[17] K. Kira, L. A. Rendell, A Practical Approach To Feature-Selection. 9th International Workshop on Machine Learning, pp. 249-256, Aberdeen, Scotland, 1992.

[18] S. Sevindik, Diskriminant analizi ve bazı alternatif regresyon analizleri. Yüksek Lisans Tezi, Çukurova Üniversitesi Fen Bilimleri Enstitüsü, Türkiye, 2018.

[19] M. Belgiu and L. Dragut, Random forest in remote sensing: A review of applications and future directions. ISPRS Journal of Photogrammetry and Remote Sensing, 114, 24-31, 2016. https://doi.org/10.1016/j.isprsjprs.2016.01.011.

[20] O. Kramer, K-Nearest Neighbors. in Dimensionality Reduction with Unsupervised Nearest Neighbors. O. Kramer, Ed. Berlin, Heidelberg: Springer, pp. 13-23, 2013.

[21] M. Awad and R. Khanna, Support Vector Machines for Classification. in Efficient Learning Machines: Theories, Concepts, and Applications for Engineers and System Designers. M. Awad and R. Khanna, Eds. Apress Berkeley, CA, pp. 39-66, 2015.

[22] M. Robnik-Sikonja and I. Kononenko, Theoretical and empirical analysis of ReliefF and RReliefF. Machine Learning, 53 (1-2), 23-69, 2003. https://doi.org/10.1023/a:1025667309714.

[23] K. Kira and L. A. Rendell, The Feature Selection Problem: Traditional Methods and a New Algorithm. In Proceedings of the 10th AAAI Conference on Artificial Intelligence, pp. 129–134, California, ABD, July 12-16, 1992.

[24] S. Riyanto, I. S. Sitanggang, T. Djatna and T. D. Atikah, Comparative Analysis using Various Performance Metrics in Imbalanced Data for Multi-class Text Classification. International Journal of Advanced Computer Science and Applications, 14 (6), 1082-1090, 2023.

[25] L. Cuadros-Rodríguez, E. Pérez-Castaño and C. Ruiz-Samblás, Quality performance metrics in multivariate classification methods for qualitative analysis. TrAC Trends in Analytical Chemistry, 80, 612-624, 2016. https://doi.org/10.1016/j.trac.2016.04.021.

[26] M. A. Günen, U. H. Atasever, E. Besdok, Analyzing the Contribution of Training Algorithms on Deep Neural Networks for Hyperspectral Image Classification. Photogrammetric Engineering and Remote Sensing 86 (9): 581-588, 2020.