Gazi Üniv. Müh. Mim. Fak. Der.
Cilt 30, No 4, 597-603, 2015

Journal of the Faculty of Engineering and Architecture of Gazi University
Vol 30, No 4, 597-603, 2015

ÖZNİTELİKLER ARASI KORELASYONUN DÜŞÜK OLDUĞU VERİ KÜMELERİNDE SINIFLANDIRMA BAŞARISINI ARTIRMAK İÇİN YOĞUNLUK TEMELLİ ÖZNİTELİK OLUŞTURMA

Selahaddin Batuhan AKBEN[1], Ahmet ALKAN[2]
[1]Osmaniye Korkut Ata Üniversitesi, Bahçe Meslek Yüksek Okulu, Osmaniye
[2]Kahramanmaraş Sütçü İmam Üniversitesi, Elektrik-Elektronik Mühendisliği Bölümü, Kahramanmaraş
batuhanakben@osmaniye.edu.tr, aalkan@ksu.edu.tr

(Geliş/Received: 12.01.2015; Kabul/Accepted: 28.07.2015)

ÖZET

Veri kümesinde, sınıfları aynı olan öznitelikler (özellikler) arasındaki korelasyon düşük ise sınıflandırma yöntemlerinin başarı oranı da düşük olmaktadır. Bu çalışmanın amacı bu tip veri kümelerinde sınıflandırıcıların başarı oranını yükseltmektir. Çalışmada veri kümesinin özniteliklerindeki değerler, öncelikle yoğunluk katsayılarına dönüştürülmüştür. Böylece öznitelikler arasında daha yüksek korelasyon bulunan yeni veri kümeleri oluşturulmuştur. Ardından bu yeni veri kümesinin asıl veri kümesinin yapısına göre sınıflandırmaya ne kadar katkıda bulunduğu değerlendirilmiştir. Değerlendirme işlemi için hem gerçek veri kümelerine hem de önerilen yöntem sayesinde oluşturulan yeni veri kümelerine çeşitli sınıflandırma yöntemleri uygulanmıştır. Sınıflandırma sonuçları karşılaştırıldığında, önerilen yöntemin sınıflandırıcıların başarı oranına yaklaşık % 17 oranda katkıda bulunduğu gözlemlenmiştir.

Anahtar Kelimeler: Öznitelik oluşturma, sınıflandırma, parzen pencereleme, yoğunluk katsayıları

DENSITY-BASED FEATURE EXTRACTION TO IMPROVE THE CLASSIFICATION PERFORMANCE IN THE DATASETS HAVING LOW CORRELATION BETWEEN ATTRIBUTES

ABSTRACT

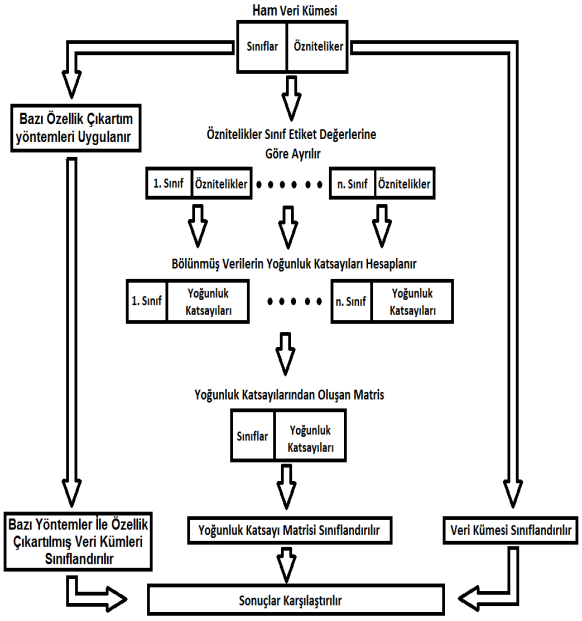
If there is low correlation between attributes belonging to the same classes of datasets, the success rates of classification methods will be low. The aim of this study is to increase the success rates of classifiers in such datasets. In this study, attributes of datasets were firstly converted to density coefficients. Thus, the new datasets having higher correlation between the attributes have been created. Then compared to the structure of the original dataset, this new dataset was evaluated in terms of contribution to classification performance. For the evaluation process, various classification methods were applied to original datasets as well as new datasets generated by the proposed method. According to comparison results, it was observed that the proposed method contributes to the classifier performance about 17%.

Keywords: Feature extraction, classification, parzen window, density coefficients

1. GİRİŞ (INTRODUCTION)

Aynı sınıfa ait farklı özniteliklerdeki veri çiftleri arasındaki korelasyon çok düşük ise öznitelik uzayındaki verilerin dağılımının rastgele olduğu bilinmektedir. Rastgele dağılmış verileri ise tek bir sınıf olarak tanımlamak zordur. Çünkü bu rastgele dağılmış veriler diğer sınıfların verileri ile iç içe geçebilir. Sonuç olarak sınıflandırma performansı bu gibi durumlarda azalır [1, 2, 3]. Sınıflandırmanın bu sorununu çözebilmek için, veri kümelerinin özniteliklerindeki verilerin başka bir değere dönüştürülmesi gerekir. Bir başka deyişle özellik çıkartımı yapılmalıdır. [4, 5, 6]. Literatürde birçok öznitelik oluşturma (özellik çıkartım) yöntemi mevcuttur. Fakat bu yöntemlerin rastgele dağılmış

veri gruplarına katkısı sınırlıdır [7, 8, 9]. Bu nedenle hala yeni bir öznitellik çıkartım yöntemine ihtiyaç vardır. Bu ihtiyacı karşılamak için özniteliklerdeki verilerin yoğunluk katsayılarına dönüştürülmesi uygun bir çözüm olabilir [10]. Çünkü sınıflandırmada amaç birlikte hareketi tespit etmektir ve veri gruplarının değer olarak yoğunlaştığı ortak noktaları belirlemek birlikte hareketin tespiti için önemli olabilir. Aynı sınıfa ait öznitelik grupları genelde birbirinden farklı veriler içerir. Özniteliklerdeki bu verilerin ortalamadan uzaklık miktarının (sınıflara ait öznitelik gruplarının standart sapma değerlerinin), büyük olması öznitelik grupları arasındaki korelasyonu olumsuz etkiler ve aynı sınıfa ait öznitelik grupları arasındaki korelasyon katsayısı ortalamasını düşürür [11]. Eğer aynı sınıfa ait öznitelik gruplarındaki verilerin yoğunluk katsayıları hesaplanırsa, yeni öznitelik gruplarının (Yoğunluk katsayılarından oluşan öznitelik gruplarının) standart sapma değerinin daha düşük olduğu bilinmektedir [12]. Böylece aynı sınıflara ait öznitelikler arasındaki korelasyon değeri de daha yüksek olmaktadır. Bu nedenle, bu çalışmada gerçek veri kümelerindeki aynı sınıfa ait özniteliklerin verileri Parzen Pencereleme Yöntemi kullanılarak yoğunluk katsayılarına dönüştürülmüştür. Sonra, yoğunluk katsayılarından oluşan yeni matris farklı sınıflandırma yöntemleri ile sınıflandırılmıştır. Bu arada asıl (ham) veri kümeleri ve bazı özellik çıkartım yöntemleri uygulanmış veriler (bazı özellik çıkartım uygulanmış ham ver veriler) de aynı sınıflandırma yöntemleri ile sınıflandırılmıştır. Son olarak, asıl veri kümelerinin sınıflandırma sonuçları ve yoğunluk katsayılarından oluşan yeni veri kümelerinin sınıflandırma sonuçları karşılaştırılmıştır. Karşılaştırma sonucunda önerilen öznitelik çıkartma yönteminin sınıflandırıcı performansına önemli oranda katkıda bulunduğu tespit edilmiştir. Bu çalışmanın akış diyagramı Şekil 1'den de görülebilir.



Şekil 1. Çalışmanın akış diyagramı (Flow diagram of this study)

2. ÇALIŞMADA KULLANILAN VERİ KÜMELERİ (DATASETS USED IN THIS STUDY)

Bu çalışmada Irvine California Üniversitesi (UCI) makine öğrenmesi veri tabanından alınan gerçek veri kümeleri (Statlog-Heart, Liver, Wine, User-Knowledge, Seeds ve Iris Data) kullanılmıştır [13]. Bu veri kümelerinin özellikleri, anılan sıraya göre aynı sınıfa ait öznitelikler arası korelasyonun artmasıdır.

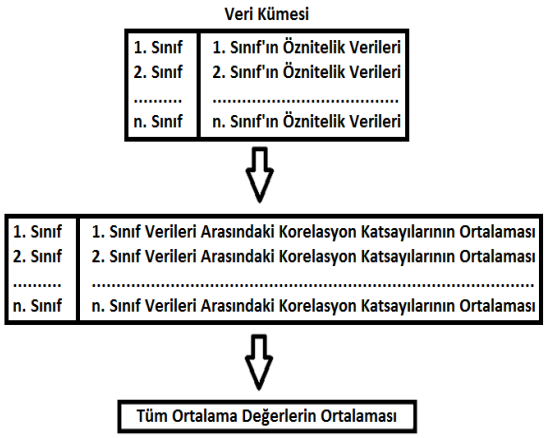
Aynı sınıfa ait öznitelikler arası korelasyon ortalaması Tablo 1'den görülebilir.

Tablo 1. Öznitelikler Arası Ortalama Korelasyon Katsayıları (Mean Correlation Coefficients Between Attributes)

Veri Kümeleri	Ortalama Korelasyon Katsayıları
Heart	0,044
Liver	0,048
Wine	0,061
User Knowledge	0,063
Seeds	0,181
Iris	0,252

Tablo 1'deki değerler şu şekilde hesaplanmıştır: Öncelikle aynı sınıfa ait özniteliklerdeki veriler arası korelasyon katsayıları tek tek hesaplanmıştır. Sonra bu korelasyon katsayılarının ortalaması bulunmuştur. Bir diğer ifade ile, her test elemanının (örneğin) öznitelikleri arasındaki korelasyon katsayılarının ortalaması bulunmuştur.

Böylece veri kümeleri hakkında korelasyon ile ilgili bilgi edinilmiştir. Veri kümeleri hakkındaki bu korelasyon bilgisinin nasıl elde edildiği ayrıca Şekil 2'den de görülebilir.



Şekil 2. Veri kümelerinin ortalama korelasyon katsayılarının elde edilişi (Obtaining of mean correlation coefficients of datasets)

Ayrıca kullanılan veri kümelerinin karakteristik özellikleri de Tablo 2'den görülebilir.

Tablo 2. Veri Kümelerinin Özellikleri (Features of Datasets)

Veri Kümeleri	Kümelerdeki Veri Sayısı	Öznitelik Vektörü Sayısı	Sınıf Sayısı
Heart	270	14	2
Liver	345	6	2
Wine	178	13	3
User Knowledge	403	5	4
Seeds	210	7	3
Iris	150	4	3

3. YÖNTEM (METHOD)

Çalışmada kullanılan veriler, yoğunluk katsayılarına Parzen Pencereleme Yöntemi kullanılarak dönüştürülmüştür. Bu yöntemde, sabit bir pencere boyutu seçilir. Sonra yoğunluk katsayısı hesaplanacak veri, pencerenin tam ortasında olacak şekilde konumlandırılır. Son olarak pencere içinde kalan veriler belirli bir matematiksel işleme göre yoğunluk katsayısını oluşturur. Bu matematiksel işlemler ise aşağıdaki denklemlerdeki gibidir. Yoğunluk Fonksiyonu $P(x)$, Denklem 1'deki gibi ifade edildiğine göre:

$$P(x) = \frac{k/n}{V} \qquad k = \sum_{i=1}^{i=n} \varphi\left(\frac{x-x_i}{h}\right) \tag{1}$$

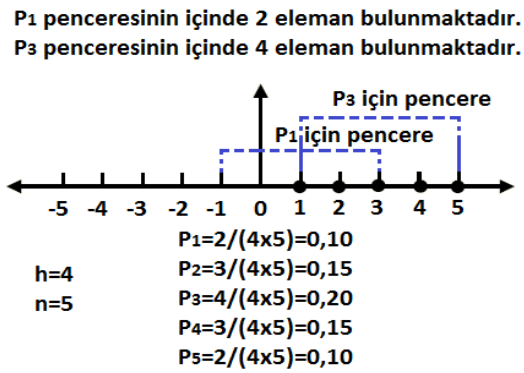
Bu Denklem 1'deki x bir örnek, n örneklerin sayısı, h pencere genişliği ve V ise hacimdir (pencere boyutu). Denklem 2'deki verilerin yoğunluk katsayılarını oluşturan bağıntı, bu yoğunluk fonksiyonunda yerine koyulursa, yoğunluk fonksiyonu denklem 3'deki gibi belirlenebilir.

$$\varphi(u) = \begin{cases} 1, & |u_j| \le \frac{1}{2},\ J = 1, \dots, d \\ 0, & Aksi\ Halde \end{cases} \tag{2}$$

$$P_\varphi(x) = \frac{1}{n}\sum_{i=1}^{i=n} \frac{1}{h^d} \varphi\left(\frac{x-x_i}{h}\right) \tag{3}$$

Burada "d" veri kümesinin boyutudur. [14, 15, 16]. Ancak, amaç her bir veri için yoğunluk katsayısı oluşturmak olduğundan yoğunluk fonksiyonunun değerlerini yalnızca verilerin bulunduğu noktalar için hesaplamak yeterlidir. Yani ayrık bir fonksiyonda verilerin bulunmadığı noktalarda fonksiyon değerini hesaplamaya gerek yoktur. Öyle ise her bir veri için pencere içinde kalan diğer elemanların sayılarının toplamının $\frac{1}{n.h^d}$ değeri ile çarpımı yoğunluk katsayısını oluşturur. Parzen Pencereleme Yöntemi ile yoğunluk katsayısı oluştururken eğer pencere büyüklüğü çok düşük olursa katsayıların birçoğu sıfır olur ve verilerin yoğunluk katsayıları tepelerden (piklerden) oluşan ayrık bir fonksiyona benzer. Birçok noktası sıfır olan ayrık fonksiyonlar ise sınıflandırmada yanıltıcı olabilir. Pencere büyüklüğü çok büyük olursa bu kez birçok katsayı birbiriyle aynı olur ve bu durum da yine sınıflandırma için uygun değildir [17].

Öyle ise pencere büyüklüğü ayarlanırken, mümkün olduğu kadar katsayıları birbirinden farklı ve sıfır olmayacak şekilde üretmek amaçlanmalıdır. Yani pencere büyüklüğü, veri grubunun yoğunlaştığı bölgelerdeki (veri grubunun merkezindeki) elemanların katsayısı büyük, merkeze uzak elemanların katsayısı ise düşük olacak şekilde ayarlanmalıdır. Eğer uzayda saçılmış bir veri grubunun birbirine en uzak iki verisi arasındaki mesafe pencere boyutu olarak seçilirse istenilen şartlar sağlanmış olur. Çünkü bu durumda merkeze uzak verilerin katsayılarının oluşumuna yaklaşık olarak kendileri ile merkez arasındaki mesafe içinde kalan elemanlar katkıda bulunur. Merkezdeki elemanların katsayılarının oluşumuna ise hemen hemen tüm elemanlar katkıda bulunmuş olur. Şekil 3'de $x = [1,2,3,4,5]$ olan örnek bir veri kümesi için pencere büyüklüğü seçimi ve sonuçları gösterilmiştir. Bu veri kümesinde birbirine en uzak elemanlar arası mesafe 4 olduğundan pencere büyüklüğü de 4 olarak seçilmiştir.



Şekil 3. Örnek bir veri kümesinde pencere büyüklüğü seçimi (Window size selection in an example dataset)

Böylece özniteliklerin yoğunluk katsayıları veri grubunun merkezine doğru artıp tekrar azalan normal dağılıma benzer ayrık bir fonksiyona dönüşür. İki farklı normal dağılım fonksiyonu arasındaki korelasyonun ise yüksek olması beklenir [18]. Çünkü fonksiyonların artış ve azalış noktaları birbirine benzediğinden birlikte hareket söz konusu olur. Üstelik veri gruplarında veriler genelde merkeze etrafında toplandığı için bu tip bir fonksiyonda veriler arası standart sapma da asıl veri grubuna göre daha düşük olur. Böylece veri gruplarının uzaydaki saçılım aralığı (rastgeleliği) azalmış olur. Şekil 3'deki örnek veri grubunun standart sapma/ortalama değeri 0,527 iken yoğunluk katsayılarının standart sapma/ortalama değeri 0,298'dir. Veri gruplarının standart sapma

değerinin düşmesi aralarındaki korelasyonun da yüksek olması anlamına gelir. Sonuçta aralarında yüksek korelasyon olan veri gruplarının sınıflandırma başarısı da artar. Bu nedenle çalışmada önerilen yöntemde pencere büyüklüğü veri gruplarının birbirine en uzak elemanları arasındaki mesafe olarak seçilmiştir. Yoğunluk katsayısı dönüşümünün özniteliklerin sınıflandırılmasına katkısı daha iyi görebilmek için gerçek bir veri grubunda inceleme yapılacak olursa: Örneğin şarap veri kümesinin (Wine) 10. ve 11. özniteliklerinin 1. sınıfa ait verilerinin standart-sapma/ortalama oranı 0.22 iken bu verilerin ortalama korelasyon değeri de 0,21'dir. Fakat verilerin yoğunluk katsayılarına ait standart-sapma/ortalama oranı 0,17 iken korelasyon değeri 0,43'dür. Bu karşılaştırmada standart-sapma/ortalama oranı kullanılmasının sebebi verilerin aynı sınır içinde değerlendirilmesidir. Bu özniteliklerin standart-sapma/ortalama oranları ve korelasyon değerleri Tablo 3'den de görülebilir.

Tablo 3. Şarap Veri Kümesinin 10. ve 11. Özniteliklerinin 1. Sınıfa Ait Verilerin Özellikleri (Features of 10. and 11. Attributes belonging to 1. Class of Wine Dataset)

Veri	Standart-Sapma/Ortalama	Korelasyon
10. ve 11. Özniteliklerin 1. Sınıfa ait Verileri	0,22	0,21
Verilerin Yoğunluk Katsayıları	0,17	0,43

Tablo 3'den de görüldüğü gibi yoğunluk katsayılarından oluşan grupların birlikte hareketi asıl veri gruplarına göre daha fazladır. Yani Yoğunluk katsayılarından oluşan grupların verileri arasındaki korelasyon asıl veri gruplarına göre daha yüksektir. Öyle ise sınıflara ait her bir öznitelik gurubu için ilk olarak birbirine en uzak elemanlar arasındaki mesafe pencere boyutu olarak belirlenmelidir. Bu işlem aşağıdaki gibi gerçekleştirilir.

$A = m \times n$ boyutlu bir veri kümesinde, veri kümesinin ilk yarısı birinci sınıf, diğer yarısı ikinci sınıf iken $P(A,h)$ yoğunluk fonksiyonu, "h" ise pencere boyutu olsun. Bu durumda D, elemanların sınıftaki diğer elemanlara olan uzaklık fonksiyonu iken h, $2 \times n$ boyutlu pencere değeridir. Böylece h, denklem 4 ve denklem 5'deki gibi hesaplanabilir. Bu işlemlerde iki adet sınıf olduğundan yine 2 adet pencere değeri olmaktadır. Eğer z adet sınıf olsa idi h, z adet sınıf olurdu.

$$i = 1, 2, \dots, n \quad h(1,i) = maks\left(D\left(A\left(1{:}\frac{m}{2}, i\right)\right)\right) \tag{4}$$

$$i = 1, 2, \dots, n \quad h(2,i) = maks\left(D\left(A\left(\frac{m}{2}+1{:}m, i\right)\right)\right) \tag{5}$$

Bununla beraber özniteliklerdeki elemanlar arası uzaklıkların homojenliği bazen çok düşük olabilir. Bu durum ise yoğunluk katsayılarında yerel maksimum ve minimumların oluşmasına sebep olur. Bu yüzden yoğunluk katsayılarında standart sapma yüksek ve korelasyon düşük olabilir. Sonuçta önerilen yöntemin bazı veri kümelerinde katkısı da beklenenden düşük olabilir. Ya da veri gruplarının birlikte hareketi (aralarındaki korelasyon) zaten yüksek ise, önerilen yoğunluk katsayısı dönüşümü yöntemi sınıflandırma başarısına beklenenden daha az katkı sağlayabilir.

4. DENEYSEL BULGULAR (EXPERIMENTAL FINDINGS)

UCI veri tabanından alınan gerçek veri kümelerine uygulanan sınıflandırma yöntemlerinin başarı oranları Tablo 4'den görülebilir. Aynı zamanda sınıflandırma yöntemlerinin yoğunluk katsayılarından oluşan özellik matrislerine uygulandığında elde edilen başarı oranları da Tablo 5'den görülebilir. Tablo 4 ve Tablo 5'deki korelasyon değerleri ham verilere aittir.

Tablo 4. Sınıflandırıcıların Veri Kümelerindeki Başarı Oranları (Classifier Performance in Datasets)

Veri Kümeleri	Korelasyon Katsayıları	KNN	Naive Bayes	DVM
Heart	0,044	67,36	85,93	83,70
Liver	0,048	68,66	57,97	66,09
Wine	0,061	75,28	98,88	96,19
User_Knowledge	0,063	83,33	90,30	67,83
Seeds	0,181	90,72	90,95	91,90
Iris	0,252	97,21	96,00	73,33

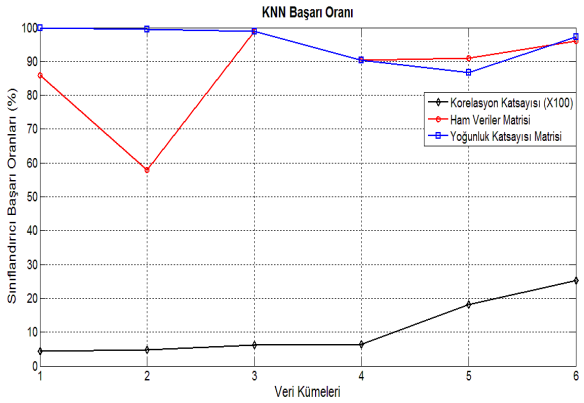
Tablo 4'den görüldüğü gibi sınıflandırıcıların başarı oranları genelde korelasyon katsayılarına orantılı olarak artmaktadır.

Tablo 5. Sınıflandırıcıların Öznitelik Matrislerindeki Başarı Oranları (Classifier Performance in Extracted Features)

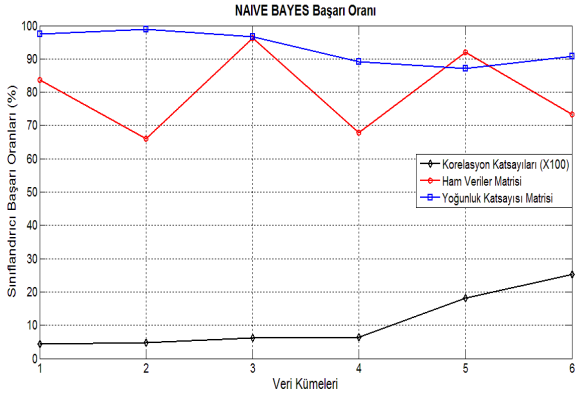
Veri Kümeleri	Korelasyon Katsayıları	KNN	Naive Bayes	DVM
Heart	0,044	100,00	100,00	97,41
Liver	0,048	99,71	99,42	98,84
Wine	0,061	98,88	98,88	96,63
User Knowledge	0,063	92,25	90,31	89,15
Seeds	0,181	87,14	86,67	87,14
Iris	0,252	94,00	97,33	90,67

Fakat Tablo 5'den görüldüğü gibi önerilen özellik çıkartım yöntemi düşük korelasyon katsayısına sahip veri kümelerinde başarıyı artırmıştır. Bununla beraber önerilen yöntem korelasyon katsayısı daha yüksek olan veri kümelerinde başarıya pek fazla katkıda bulunamamıştır.

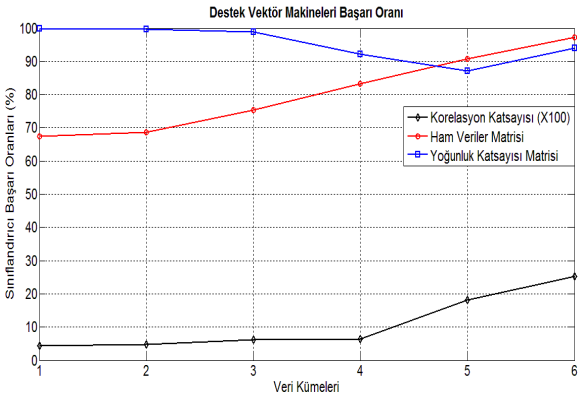
Ayrıca Tablo 4 ve Tablo 5'de görülen korelasyon katsayıları ve sınıflandırıcı başarı oranları arasındaki ilişki Şekil 4, Şekil 5 ve Şekil 6'dan da görülebilir.



Şekil 4. KNN yönteminin başarı oranı ile yoğunluk temelli öznitelik çıkartım yöntemi arasındaki ilişki (Relation between classifier performance of KNN method and density-based feature extraction method)



Şekil 5. Naive Bayes yönteminin başarı oranı ile yoğunluk temelli öznitelik çıkartım yöntemi arasındaki ilişki (Relation between classifier performance of Naïve Bayes method and density-based feature extraction method)



Şekil 6. DVM yönteminin başarı oranı ile yoğunluk temelli öznitelik çıkartım yöntemi arasındaki ilişki (Relation between classifier performance of DVM method and density-based feature extraction method)

Şekil 4, Şekil 5 ve Şekil 6'dan görüldüğü gibi veri kümelerinin korelasyon katsayısı arttıkça ham veriler için sınıflandırıcıların başarı oranları da artmaktadır. Yoğunluk katsayıları için ise durum farklıdır. Yoğunluk katsayıları dönüşümü düşük korelasyonlu veri kümelerinde sınıflandırıcı başarısını ham verilere göre ortalama %18 oranda artırmaktadır. Fakat öznitelikler arası korelasyonu yüksek olan veri kümelerinde yoğunluk katsayısı dönüşümü çok fazla işe yaramamaktadır. Bununla beraber önerilen yöntem sayesinde sınıflandırıcı başarıları korelasyondan daha bağımsız ve kararlı hale gelmiştir. Literatürde en çok kullanılan özellik çıkartım yöntemi Wraper Yöntemi olup alternatif olarak Embedded yöntemi de yine literatürde alternatif olarak önerilmiştir [7,8,9,19]. Bu çalışmada önerilen yöntemin, öznitelikler arası korelasyonun düşük olduğu veri kümlerinde, literatürde önerilen diğer özellik çıkartım yöntemlerine göre başarısı ise Tablo 6 ve Tablo 7'nin Tablo 5 ile kıyaslanması sonucu görülebilir. Böylece bu çalışmada önerilen yöntemin diğer özellik çıkartım yöntemlerine göre başarısı da değerlendirilebilir.

Tablo 6. Sınıflandırıcıların Wrapper özellik çıkartım yöntemi uygulanmış veri kümelerindeki başarı oranları (Classifier Performance in Extracted Features by Wrapper Method)

Veri Kümeleri	Korelasyon Katsayıları	KNN	Naive Bayes	DVM
Heart	0,044	68,32	86,11	84,26
Liver	0,048	69,52	58,07	66,56
Wine	0,061	75,34	98,88	96,23
User Knowledge	0,063	82,67	89,03	67,98
Seeds	0,181	91,12	92,25	92,10
Iris	0,252	98,13	96,21	73,66

Tablo 7. Sınıflandırıcıların Embedded özellik çıkartım yöntemi uygulanmış veri kümelerindeki başarı oranları (Classifier Performance in Extracted Features by Embedded Method)

Veri Kümeleri	Korelasyon Katsayıları	KNN	Naive Bayes	DVM
Heart	0,044	68,66	87,23	85,06
Liver	0,048	69,78	59,17	67,04
Wine	0,061	76,38	98,88	97,01
User Knowledge	0,063	83,33	90,30	67,83
Seeds	0,181	91,21	91,65	92,30
Iris	0,252	97,91	96,15	74,92

Tablo 6 ve Tablo 7'deki başarı oranları, Tablo 5'deki önerilen yöntemin başarı oranları ile kıyaslandığında, önerilen yöntemin öznitelikler arası korelasyonun düşük olduğu veri kümlerinde mevcut özellik çıkartma yöntemlerine göre de daha başarılı olduğu görülmektedir. Tablo 5 ile Tablo 6 kıyaslandığında bu çalışmada önerilen yöntemin, Wrapper Yöntemi'ne göre ortalama %17,6 daha başarılı olduğu görülmektedir. Tablo 5 ile Tablo 7 kıyaslandığında ise bu çalışmada önerilen yöntemin, Embedded Yöntemi'ne göre ortalama %17,1 daha başarılı olduğu görülmektedir. Bir diğer ifade ile literatürdeki bu iki özellik çıkartım yönteminin başarısı, öznitelikler arası korelasyonun düşük olduğu veri kümelerinde ham verilerin sınıflandırıcı başarı oranına göre pek de farklı değildir.

5. SONUÇLAR VE TARTIŞMA (RESULTS AND DISCUSSIONS)

Veri kümelerindeki öznitelikler arası korelasyon sınıflandırma başarısını etkileyen önemli bir etkendir. Eğer aynı sınıfa ait öznitelik verileri arasındaki korelasyon düşük olursa sınıflandırma başarısı da düşük olur. Bu çalışmada önerilen yöntem, aynı sınıfa ait öznitelik verileri arasındaki korelasyon katsayılarının düşük olduğu veri kümelerinde sınıflandırma başarısını artırmaktadır. Önerilen yöntem, özniteliklerin standart sapma değerlerini düşürüp aralarındaki korelasyonu yükselterek sınıflandırma başarısını yükseltmeyi amaçlar. Bunu gerçekleştirebilmek için de yoğunluk temelli bir özellik çıkartım yöntemi kullanmaktadır. Böylece yoğunluk katsayılarından oluşan özellik matrisinin veriler arası korelasyon değerleri asıl veri kümelerine göre daha yüksek olmaktadır. Bununla beraber önerilen yöntem, veriler arası korelasyon değerleri zaten yüksek olan veri kümelerinde pek fazla işe yaramamaktadır.

Ayrıca önerilen yöntemin sınıflandırma başarısına katkısı korelasyon ile doğrusal orantılı da değildir. Çünkü sınıflandırma başarısı yalnızca korelasyona bağlı değildir. Bu nedenle önerilen yöntem sınıflandırma için veri yapısının diğer şartlara bağlı olarak uygun olmadığı veri kümelerinde çok fazla katkı sağlamayabilir. Ancak yalnızca veriler arası korelasyona bağlı olarak sınıflandırma başarısı düşük olan veri kümelerinde çok kullanışlıdır. Bir başka açıdan bakıldığında önerilen yöntem sayesinde sınıflandırma başarıları daha kararlı ve sabit bir hale gelmiştir. Böylece sınıflandırma başarısı ham veriler arası korelasyondan daha bağımsız hale gelmiştir. Bu sonuç ise çeşitli sebeplerden dolayı sınıflandırma başarısı düşük olan deneklerin de daha rahat değerlendirilmesini sağlamaktadır. Özellikle dış etkenlerden çok etkilenen biyolojik verilerin sınıflandırılması, önerilen yöntem sayesinde daha kolay olabilir.

KAYNAKLAR (REFERENCES)

1. Kenzie, G.M. ve Peng, D., **Statistical Modelling in Biostatistics and Bioinformatics Contributions to Statistics**, Springer, 2014.
2. Köklü, M., Kahramanlı, H., Allahverdi, N., "Sınıflandırma Kurallarının Çıkarımı İçin Etkin ve Hassas Yeni Bir Yaklaşım", **Journal of the Faculty of Engineering and Architecture of Gazi University,** Cilt 29, No 3, 477-486, 2014.
3. Hsu, H.H., Hsieh, C.W., "Feature Selection via Correlation Coefficient Clustering", **Journal of software**, Cilt 5, No 12, 1371-1377, 2010.
4. Deisy, C., Subbulakshmi, B., Baskar, S., Ramaraj, N., "Efficient Dimensionality Reduction Approaches for Feature Selection," **International Conference on Computational Intelligence and Multimedia Applications**, 121-127, 2007.
5. Guyon, I., Nikravesh, M., Gunn, S., Zadeh, L.A., **Feature Extraction Foundations and Applications**, Springer Berlin Heidelberg, 2006.
6. Nixon, M., Aguado, A., **Feature Extraction and Image Processing**, Academic Press is an Imprint of Elsevier, 2008.
7. Kohavi, R., John, G., "Wrappers for Feature Subset Selection," **Artificial Intelligence**, Cilt 97, 273-324, 1997.
8. Quinlan, J.R., **Discovering Rules from Large Collections of Examples: A Case Study, In Michie, D. ed.**, Expert Systems in the Microelectronic Age, Scotland: Edinburgh University Press, Edinburgh, 1979.
9. Liu, Y., Yin, Y.F., Gao, J.J., Tan, C. G., "Wrapper Feature Selection Optimized SVM Model for Demand Forecasting," **The International Conference on Young Computer Scientists**, 953-958, 2008.
10. Qian, L., Honggang, Q., Jun, M, Wentao, Z, Guiping, S., "One-Class Classification with Extreme Learning Machine," **Mathematical Problems in Engineering**, Article ID 412957, in press.
11. Rodgers, J.L., Nicewander, W.A., "Thirteen ways to look at the correlation coefficient", **The American Statistician**, Cilt 42, No 1, 59–66, 1988.
12. Park, S., Bera, Y. Anil, K. "Maximum Entropy Autoregressive Conditional Heteroskedasticity Model", **Journal of Econometrics (Elsevier),** Cilt 150, No 2, 219–230, 2009.
13. Frank, A., Asuncion, A., UCI Machine Learning Repository., 2014, http://www.archive.ics.uci.edu/ml
14. Babich, G.A., "Weighted Parzen windows for pattern classification", **IEEE Transactions on Pattern Analysis and Machine Intelligence**, Cilt 18, No 5, 567-570, 2002.
15. Wanga, S., Chungb, F., Xionga, F., "A novel image thresholding method based on Parzen

window estimate", **Pattern Recognition**, Cilt 41, No 1, 117-129, 2008.

16. Erdogmus, D., Hild, K.E., Principe, J.C., Lazaro, M., Santamaria, I., "Adaptive Blind Deconvolution of Linear Channels Using Renyi's Entropy with Parzen Window Estimation", **IEEE Transactions on Signal Processing**, Cilt 52, No 6, 1489-1498, 2004.

17. Veon, K.L. "Dept. of Electr. & Comput. Localized support vector machines using Parzen window for incomplete sets of categories", **Applications of Computer Vision (WACV), IEEE Workshop on**, 448-454, 2011.

18. Yuichi T., Paul, J. Choi, G.L., Huiyi C., Mohan, B., Jeremy, H., Andrew, E.X., Sunney, X, "Quantifying E. coli Proteome and Transcriptome with Single-Molecule Sensitivity in Single Cells", **Science**, Cilt 329, No 5991, 533-538, 2010.

19. Saeys, Y., Inza, I., Larranaga, P., "A review of feature selection techniques in bioinformatics", **Bioinformatics**, Cilt 23, No 19, 2507-2517, 2007.