Yayın Geliş Tarihi (Submitted): **05/04/2021**

Yayın Kabul Tarihi (Accepted): **22/06/2021**

Makele Türü (Paper Type): **Araştırma Makalesi – Research Paper**

Please Cite As/Atıf için:
Erbayram T. ve Erişoğlu M. (2021), Yeni bir özellik seçim yöntemi ve özellik seçim yöntemlerinin sınıflama performanslarının karşılaştırılması, *Nicel Bilimler Dergisi*, 3(1), 72-90. *doi: 10.51541/nicel.909876*

YENİ BİR ÖZELLİK SEÇİM YÖNTEMİ VE ÖZELLİK SEÇİM YÖNTEMLERİNİN SINIFLAMA PERFORMANSLARININ KARŞILAŞTIRILMASI

Tenzile Erbayram[1] ve Murat Erişoğlu[2]

ÖZET

Özellik seçimi, veri analizinde veri hazırlamak için uygulanan ön işlemlerden biridir. Özellik seçimi basitçe orijinal özellik kümesinden en uygun özelliklerin alt kümesinin seçim işlemidir. Bu yöntemler, orijinal veri setinde alakasız ve gereksiz bilgiyi belirlemeye ve kaldırmaya çalışır. Bu çalışmada sınıf bilgisi kullanılarak değişim katsayısına dayalı yeni bir özellik seçim yöntemi önerilmiştir. Önerilen özellik seçim yönteminin etkinliği, gerçek veri setleri kullanılarak diğer iyi bilinen özellik seçim yöntemleri ile karşılaştırılarak değerlendirilmiştir. Özellik seçim yöntemlerinin performansı, karesel diskriminant analizinde sınıflama doğruluğu ve entropi kriterleri bakımından incelenmiştir. Çalışmada birim sayısının özellik sayısından fazla olduğu nicel verilerden oluşan üç gerçek veri seti kullanılmıştır. Her bir özellik seçim yöntemine göre önem sırası belirlenen özelliklerinden ilk d adet özellik kullanılarak karesel diskriminant analizi gerçekleştirilmiştir. Özellik sayısına göre özellik seçim yöntemlerinin karesel diskriminant analizindeki sınıflama doğruluğu ve entropi değerleri hesaplanmıştır. Çalışma sonuçları, önerilen

Bu çalışma Prof. Dr. Murat ERİŞOĞLU danışmanlığında Necmettin Erbakan Üniversitesi Fen Bilimleri Enstitüsü İstatistik Anabilim Dalında Tenzile ERBAYRAM tarafından hazırlanan "Boyut İndirgeme Tekniklerinin Sınıflandırma Performanslarının Karşılaştırılması" isimli Yüksek Lisans tezinden oluşturulmuştur.

[1] Sorumlu Yazar, Arş. Gör., Selçuk Üniversitesi, Fen Fakültesi, İstatistik Bölümü, Konya, Türkiye, ORCID ID: https://orcid.org/0000-0002-3275-120X

[2] Prof. Dr., Necmettin Erbakan Üniversitesi, Fen Fakültesi, İstatistik Bölümü, Konya, Türkiye, ORCID ID: https://orcid.org/0000-0002-4589-1383

özellik seçim yönteminin hesaplama basitliği ve etkinlik açısından sınıflama analizleri için iyi bilinen diğer özellik seçim yöntemleri karşısında güçlü bir alternatif olduğunu ortaya koymuştur.

Anahtar Kelimeler: Özellik Seçimi, Entropi, Sınıflama Doğruluğu, Karesel Diskriminant Analizi

A NEW FEATURE SELECTION METHOD AND COMPARISON OF CLASSIFICATION PERFORMANCES OF FEATURE SELECTION METHODS

ABSTRACT

Feature selection is one of the preliminary processes applied to data preparation in data analysis. Feature selection is simply the process of selecting the most suitable subset of features from the original feature set. These methods try to identify and remove irrelevant and unnecessary information from the original data set. In this study, a new feature selection method based on the coefficient of variation using class information is proposed. The efficiency of the proposed feature selection method has been evaluated by comparing it with other well-known feature selection methods using real data sets. The performance of feature selection methods was examined in terms of classification accuracy and entropy criteria in quadratic discriminant analysis. Three real data sets consisting of quantitative data with the number of units less than the number of features were used in the study. Quadratic discriminant analysis was carried out using the first d features of the features whose importance order was determined according to each feature selection method. Classification accuracy and entropy values in quadratic discriminant analysis of feature selection methods were calculated according to the number of features. The results of the study revealed that the proposed feature selection method is a strong alternative to other well-known feature selection methods for classification analysis in terms of computational simplicity and efficiency.

Keywords: Feature Selection, Entropy, Classification Accuracy, Quadratic Discriminant Analysis

1. GİRİŞ

Küreselleşme ve hızla gelişen teknolojinin yaşamımızdaki etkileri iyice belirgin hale gelmiştir. Bu etki yaşam alanımızda yer alan olguların etkileşimini de arttırdığından

gerçekleştirilen bilimsel çalışmaların çoğunda göz önüne alınması gereken özellik sayısını da arttırmıştır. Özellik sayısının fazla olması ve özellikler arasında etkileşimlerin olması, veri setlerinin analizinde sonuçların özetlenmesini ve yorumlanmasını güçleştirmektedir. Verinin araştırılan tüm özellik kümesini kullanmak hesaplama ve yorumlama zorluğu yanında bazen analiz için gerekli varsayımlarının sağlanamamasına bazen de aşırı uyum problemi nedeni ile oluşturulan modellerin genelleme performansının düşmesine neden olur. Çok sayıda özelliğin olduğu veri setlerinde çoğu özellik birbiri ile ilişkilidir ve bu modellin uyumunu anlamsız bir şekilde artırır. Böyle durumlarda modelin oluştuğu eğitim setlerinde model çok iyi bir performans gösterirken test verilerinde model performansı çok düşük gerçekleşir. Çünkü model veriye aşırı bağımlıdır. Bu nedenle aşırı uyumluluk durumunda model genelleme özelliğini kaybeder ve bu istenmeyen bir durumdur.

Özellik sayısının çok fazla olduğu veri kümelerinin görselleştirilebilmesi, analizlere uygun hale getirilebilmesi, veriden daha basit ve daha anlamlı modeller üretebilmesi için boyut indirgemek gerekir. Boyut indirgeme en basit anlatımla, orijinal verinin taşıdığı bilgiden mümkün olduğunca az bir kayıpla boyut sayısını azaltma sürecidir. Boyut indirgeme özellik seçimi veya özellik çıkarımına göre gerçekleştirilir. Özellik çıkarımı tüm özellikleri dikkate alarak bu özelliklerin doğrusal veya doğrusal olmayan bileşenleri ile daha az boyutta veriyi temsil etme sürecidir. Özellik seçimi veri kümesini en iyi temsil edecek bir alt özellik kümesinin seçimi olarak tanımlanabilir. Özellik seçimi, verinin temsilinde daha az önemli olan özelliklerin veri kümesinden çıkartılması ile boyut indirgeme işlemini gerçekleştirir. Özellik seçimin yöntemleri sadece istatistiksel ölçütlere dayalı olan filtreleme yöntemleri, özellikler üzerinde arama işlemleri gerçekleştiren sarmal yöntemler ve en iyi bölen ölçütünü bulmaya dayalı olan gömülü yöntemler olmak üzere genel olarak üç grupta toplanmaktadır (Saeys vd., 2007).

İstatistik temelli filtreleme özellik seçim yöntemleri, istatistik ölçütler kullanılarak her bir özellikler ile hedef değişken arasındaki ilişkinin değerlendirilmesini ve hedef değişkenle en güçlü ilişkiye sahip olan özelliklerin seçilmesini içerir. Bu yöntemler hızlı ve etkili olabilir, ancak istatistiksel ölçülerin seçimi hem girdi hem de çıktı değişkenlerinin veri türüne bağlıdır. Bu nedenle, filtre tabanlı özellik seçimini gerçekleştirirken bir veri kümesi için uygun bir istatistiksel ölçümün seçilmesi zor olabilir.

Boyut indirgeme temelinde özellik seçimi ve özellik çıkarımı ile ilgili çok sayıda çalışma mevcuttur. Singh vd. (2016), yüksek boyutlu verilerde özellik seçme yöntemleri ile ilgili literatürde yer alan önceki çalışmaları değerlendirmişlerdir. Yıldız vd. (2016), özellik çıkarım yöntemi olan doğrusal boyut indirgeme yöntemlerinin sınıflama performanslarını karşılaştırmışlardır. Castro vd. (2018), boyut indirgeme ve çoklu dizin parçalanması için yeni bir yöntem önermişlerdir. Budak (2018), özellik seçimi için yeni bir yaklaşım önermiştir. Guo vd. (2018), video sınıflandırması için özellik seçim temelli boyut indirgeme yöntemlerini karşılaştırmışlardır.

Verinin yapısı ve özelliklerin ölçümlerindeki farklılık gibi nedenlerle uygun boyut indirgeme yönteminin seçimi önemli bir problemdir. Özellik sayısının birim sayısından fazla olduğu yüksek boyutlu verilerde, istatistiksel ölçütlere dayalı özellik seçim yöntemi olan filtreme yöntemlerinin sınıflama performanslarının karşılaştırılmasının gerçekleştirileceği çalışmada, sınıf bilgisi kullanılarak değişim katsayısına dayalı olarak yeni bir özellik seçim yöntemi önerilmiştir. Önerilen özellik seçim yönteminin etkinliği, yüksek boyutlu veri setlerinde diğer iyi bilinen özellik seçim yöntemleri ile karesel diskriminant analizindeki sınıflama performansları bakımından karşılaştırılarak değerlendirilecektir. Nicel verilerden oluşan gerçek veri setlerinde gerçekleştirilecek karşılaştırmada sınıflama doğruluğu ve entropi kriterleri kullanılacaktır.

Çalışmanın ikinci bölümünde yaygın kullanıma sahip ve iyi bilinen özellik seçim yöntemleri ve önerilen özellik seçim yöntemi verilecektir. Bu bölümde ayrıca karesel diskriminant analizi ve karşılaştırma kriterleri tanımlanacaktır. Çalışmanın uygulama bölümünde gerçek veri setleri ile ilgili tanımlamalar verildikten sonra, özelik seçim yöntemlerine göre seçilen özellikler ile gerçekleştirilen karesel diskriminant analizi sonuçlarına dayalı sınıflama doğrulukları ve entropi değerleri hesaplanacaktır. Sonuç ve öneriler bölümünde çalışmadan elde edilen sonuçlar özetlenecek ve sonraki çalışmalar için önerilerde bulunulacaktır.

2. MATERYAL VE METOT

2.1. Özellik Seçim Yöntemleri

Özellik seçimi, tahmine dayalı bir model geliştirirken özellik sayısını azaltma işlemidir. Özellik seçimi, öncelikle bilgilendirici olmayan veya gereksiz tahmin unsurlarını modelden kaldırmaya odaklanır (Kuhn ve Johnson, 2013). Hem modelin hesaplanmasını basitleştirmek hem

de bazı durumlarda model performansını iyileştirmek için özellik sayısının azaltılması gerekir. Bu bölümde filtreleme tabanlı iyi bilinen özellik seçim yöntemlerinden değişim katsayısı, F testi, Kruskal Wallis testi, Fisher skoru, komşuluk bileşen analizi, Relief ve ReliefF algoritmaları hakkında bilgi verilecektir. Ayrıca sınıf bilgisi kullanılarak değişim katsayına dayalı yeni bir özellik seçim yöntemi tanımlanacaktır.

2.1.1. Değişim Katsayısı

$\boldsymbol{X}$ sürekli rassal değişkeninin ortalaması $\boldsymbol{\mu}$ ve standart sapması $\boldsymbol{\sigma}$ olsun. Bu durumda $\boldsymbol{X}$ sürekli rassal değişkeni için değişim katsayısı

$$\boldsymbol{DK = \frac{\sigma}{\mu}} \tag{2.1}$$

eşitliği ile belirlenir. Değişim katsayısı ölçü birimi içermediği ve terim büyüklüklerinden etkilenmediği için iki kitlenin değişkenliğinin karşılaştırılmasında kullanılan istatistiksel bir ölçüdür. Değişkenliğin fazla olması veri seti içerisinde farklı özellikte birimlerin olabileceğinin yani kitle içerisinde alt kitlelerin olduğunun bir göstergesi olduğundan dolayı sınıflama için değişim katsayısı büyük olan özelliklerin seçilmesi önerilir.

2.1.2. F ve Kruskall Wallis Test İstatistikleri

Bağımsız k grup ortalamasının eşitliğinin test edilmesinde kullanılan tek yönlü varyans analizindeki F test istatistiği

$$F = \frac{(n-k)\sum_{i=1}^{k} n_i(\bar{x}_i - \bar{x})^2}{(k-1)\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(x_{ij} - \bar{x}_i\right)^2} \tag{2.2}$$

eşitliği ile hesaplanır. Eşitlikte yer alan x_{ij} gösterimi i. gruptaki j. birimin gözlem değerini ifade ederken $\bar{x}_i$ gösterimi ilgili özellik için i. grup ortalamasını ve $\bar{x}$ gösterimi ise genel ortalamayı ifade etmektedir. F test istatistiğinin büyük değer alması ilgili özelliğin sınıflama performansının yüksek olduğunu gösterir.

Tek yönlü varyans analizinde kullanılan test istatistiği parametrik bir yaklaşımdır ve analiz sonuçlarının güvenirliği için normallik ve varyansların eşitliği gibi ön varsayımların sağlanması gerekir. Varsayımların sağlanamaması durumunda bağımsız k grup ortalamasının eşitliğinin test edilmesinde kullanılan parametrik olmayan test istatistiği ise Kruskal-Wallis (KW) test

istatistiğidir. KW test istatistiği hesaplanırken gözlem değerleri yerine gözlem değerlerinin sıra puanları kullanılır ve

$$H = \frac{12}{N(N+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3(n+1) \tag{2.3}$$

eşitliği ile hesaplanır. Eşitlikte yer alan R_i gösterimi i. grupta yer alan gözlemlerin sıra puanlarının toplamını göstermektedir. KW testinde de test istatistiğinin büyük değer alması ilgili özelliğin sınıflama performansının yüksek olduğu anlamına gelir.

2.1.3. Fisher Skoru

İki sınıf durumunda Fisher'in doğrusal diskriminant analizine dayalı Fisher skoru kullanılarak her bir özelliğin sınıflamadaki önemliliği belirlenebilir. İki sınıf için sınıf ortalamaları $\bar{x}_i^+$ ve $\bar{x}_i^-$, sınıflar için standart sapmalar s_i^+ ve s_i^- olmak üzere i. özellik için Fisher skoru

$$Fisher(x_i) = \frac{|\bar{x}_i^+ - \bar{x}_i^-|}{|s_i^+ + s_i^-|} \tag{2.4}$$

eşitliği ile belirlenir. Fisher skorunun büyük değer alması iki sınıfın birbirinden iyi ayırt edilebildiğini göstermektedir (Budak, 2018). Bu çalışmada Fisher skoru k sınıf için, k adet kukla sınıf etiketi ile elde edilen k adet Fisher skorunun ortalamasını kullanılarak genelleştirilmiştir.

2.1.4. Komşuluk Bileşen Analizi

Komşuluk bileşen analizi (NCA), sınıflama algoritmalarının doğru sınıflama olasılığını en büyüklemeyi amaçlayan parametrik olmayan özellik seçim algoritmasıdır. Komşuluk bileşen analizinde doğru sınıflama olasılığını en büyükleyecek şekilde özelliklerin ağırlıklandırılması gerçekleştirilir. Elde edilen ağırlıklarla özellik seçimi gerçekleştirilir.

$T = \{(\mathbf{x}_1, \mathrm{y}_1), \ldots\ldots (\mathbf{x}_\mathrm{i}, \mathrm{y}_\mathrm{i}), \ldots\ldots (\mathbf{x}_\mathrm{n}, \mathrm{y}_n)\}$ bir eğitim verisi olsun, burada $\mathbf{x}_\mathrm{i}$ p boyutlu bir değişken vektörü, $\mathrm{y}_\mathrm{i} \in \{1, 2, \ldots, \mathrm{C}\}$ ise sınıf etiketidir. Komşuluk bileşen analizinde amaç en yakın komşu sınıflama algoritmasının doğru sınıflama olasılığını optimize eden özellik alt kümesini seçecek bir ağırlık vektörü $\mathbf{w}$ bulmaktır. Ağırlık vektörü $\mathbf{w}$ olmak üzere $\mathbf{x}_\mathrm{i}$ ve $\mathbf{x}_\mathrm{j}$ birimleri arasındaki ağırlıklı uzaklık

$$d_{\mathbf{w}}(\mathbf{x}_i, \mathbf{x}_j) = \sum_{l=1}^{d} w_l^2 \left| x_{il} - x_{jl} \right| \tag{2.5}$$

eşitliği ile ifade edilir. Eşitlikte yer alan w_l gösterimi, l. özellikle ilişkili ağırlıktır. En yakın komşu sınıf algoritmasının başarılı olması için, sezgisel ve etkili bir strateji, eğitim verisi T'de doğru sınıflama olasılığını en üst düzeye çıkarmaktır. Komşuluk bileşen analizinde referans noktası bir olasılık ile belirlenir. Burada $\mathbf{x}_i$ gözlem vektörünün referans noktası olarak $\mathbf{x}_j$ gözlem vektörünü seçmesi olasılığı

$$p_{ij} = \begin{cases} \dfrac{k(d_w(x_i, x_j))}{\sum_{k \neq i} K(d_w(x_i, x_k))}, & i \neq j \\ 0, & i = j \end{cases} \tag{2.6}$$

olarak tanımlanır. Bir $k(.)$ gösterimi $d_w(x_i, x_j)$ ağırlı uzaklık fonksiyonunun büyük değerleri için küçük değerler veren Kernel ve benzeri bir fonksiyondur. Yang vd. (2012) tarafından $k(z) = e^{-\frac{z}{\sigma}}$ şeklinde önerilmiştir. Eşitlik (2.6)'da tanımlanan olasılıklara dayalı olarak x_i gözlemim vektörünün doğru sınıflandırılma olasılığı

$$p_i = \sum_j y_{ij} p_{ij} \tag{2.7}$$

şeklinde tanımlanır. Eşitlikte yer alan y_{ij} katsayısı $\mathbf{x}_i$ gözlem vektörü ile $\mathbf{x}_j$ gözlem vektörü aynı sınıfta ise yani $y_i = y_j$ ise 1, aksi takdirde 0 değerini alır. Genel doğru sınıflama olasılığı

$$F(w) = \frac{1}{n} \sum_{i=1}^{n} p_i = \frac{1}{n} \sum_{i=1}^{n} \sum_j y_{ij} p_{ij} \tag{2.8}$$

eşitliği ile ifade edilir. Özellik seçimi yapmak ve aşırı uyumu azaltmak için Eşitlik (2.8)'e bir ceza terimi ekleyerek ilgili fonksiyon

$$F(w) = \frac{1}{n} \sum_{i=1}^{n} p_i - \lambda \sum_{l=1}^{p} w_l^2 = \frac{1}{n} \sum_{i=1}^{n} \underbrace{\left[\sum_j y_{ij} p_{ij} - \lambda \sum_{l=1}^{p} w_l^2 \right]}_{F_i(w)} \tag{2.9}$$

$$= \frac{1}{n} \sum_{i=1}^{n} F_i(w)$$

şeklinde düzenlenebilir. Burada λ düzeltme parametresidir. Komşuluk bileşen analizinde Eşitlik (2.9) ile verilen amaç fonksiyonunu en büyükleyecek şekilde w ağırlık vektörü tahmin edilir.

$$\hat{\mathrm{w}} = arg\max_{\mathrm{w}}\left(\frac{1}{n}\sum\nolimits_{i=1}^{n} F_i(\mathrm{w})\right) \tag{2.10}$$

2.1.5. Relief ve ReliefF Algoritmaları

Relief algoritması, iki sınıflı veri kümelerinde özellik seçimi için Kira ve Rendell (1992) tarafından önerilmiştir. Relief algoritması özelliklerin, birimlerin ait olduğu sınıftaki en yakın komşuları ile ait olmadığı sınıftaki en yakın komşuları arasındaki uzaklık farklılıklarına dayalı olarak ağırlıklandırılması esasına dayanmaktadır.

Kira ve Rendell (1992) tarafından önerilen orijinal Relief algoritmasında tüm birimler için ait oldukları sınıftaki en yakın komşu uzaklığı $\Delta(+) = (x_i - x^+)^2$ ve ait olmadıkları sınıftaki en yakın komşu uzaklığı $\Delta(-) = (x_i - x^-)^2$ olmak üzere ilgili özelliğin ağırlığı

$$W(x) = \frac{\sum_{i=1}^{n}\{\Delta(-) - \Delta(+)\}}{n} \tag{2.11}$$

eşitliği ile hesaplanır. Eşitliklerde yer alan x^+ gösterimi x_i gözlem değerinin aynı sınıfındaki en yakın komşu gözlemini, x^- gösterimi ise x_i gözlem değerinin ait olmadığı sınıfdaki en yakın komşu gözlemini temsil etmektedir. Aynı sınıfta bulunan komşular arasındaki uzaklığın küçük, farklı sınıfta bulunan komşular arasındaki uzaklığın büyük olması beklenir. Buna göre Relief algoritmasındaki ağırlık değerinin büyük olması ilgili özelliğin sınıflama performansının yüksek olduğunu gösterir. Algoritmanın son adımında ilgili özellikler içerisinden belirlenen eşik değerini aşan özellikler seçilerek boyut azaltma işlemi gerçekleştirilir.

İki sınıf problemi için önerilen Relief algoritması Kononenko (1994) tarafından k sınıf için genelleştirilmiştir. Kononenko (1994) tarafından önerilen algoritma ReliefF olarak isimlendirilmiştir. ReliefF algoritmasında en yakın m komşu üzerinden ağırlıklar hesaplanmıştır. Aynı sınıfta bulunan en yakın komşu uzaklıkların ağırlığa katkısı

$$W(x)^r = W(x)^{r-1} - \frac{\Delta(x_i, x^+)}{m} d_{i,+} \tag{2.12}$$

eşitliği ile oluşturulur. Eşitlikte yer alan $\Delta(x_i, x^+)$ gösterimi x_i birimi ile aynı sınıfta yer alan m gözlem noktasından birini temsil eden x^+ gözlemi arasındaki uzaklığı ifade eder ve sayısal veriler için bu uzaklık

$$\Delta(x_i, x^+) = \frac{|x_i - x^+|}{\max(x) - \min(x)} \tag{2.13}$$

eşitliği ile hesaplanır. Eşitlik (2.12)'de yer alan $d_{i,+}$ gösterimi ise en yakın komşunun yakınlık derecesine göre ağırlığını ifade eder ve

$$d_{i,+} = \frac{\tilde{d}_{i,+}}{\sum_{l=1}^{m} \tilde{d}_{i,l}} \tag{2.14}$$

eşitliği ile hesaplanır. Eşitlikte yer alan $\tilde{d}_{i,l} = e^{-(rank(i,l)/sigma)^2}$ şeklinde hesaplanır. Eşitlikte yer alan $sigma$ değeri araştırmacı tarafından belirlenen pozitif bir sayıdan oluşan ölçeklendirme değeridir.

Farklı sınıfta bulunan en yakın komşu uzaklıkların ağırlığa katkısı

$$W(x)^r = W(x)^{r-1} + \frac{p_+}{1-p_-} \frac{\Delta(x_i, x^-)}{m} d_{i,-} \tag{2.15}$$

eşitliği ile hesaplanır. Eşitlikte yer alan p_+ gösterimi x_i biriminin ait olduğu sınıfa ait önsel olasılığını gösterirken, p_- gösterimi x^- biriminin ait olduğu sınıfa ait önsel olasılığı göstermektedir. İlgili önsel olasılıklar $p_+ = \frac{n_+}{n}$ ve $p_- = \frac{n_-}{n}$ eşitlikleri ile elde edilir. ReliefF algoritmasında da ağırlık değerinin büyük olması ilgili özelliğin sınıflama performansının yüksek olduğunu göstermektedir.

2.1.6. Önerilen Özellik Seçim Yöntemi

Sınıf üyeliklerinin bilindiği veri setlerinde sınıf bilgisi kullanılarak değişim katsayısının özellik seçiminde farklı bir kullanımı söz konusu olabilir. X sürekli rassal değişkeninin $\mu_1, \mu_2, \ldots, \mu_K$ ortalamaları ve $\sigma_1, \sigma_2, \ldots, \sigma_K$ standart sapmalarına sahip k tane alt kitleye sahip olduğu varsayılsın. Bu durumda i. alt kitleye ilişkin değişim katsayısı

$$DK_i = \frac{\sigma_i}{\mu_i} \qquad i = 1, 2, \cdots, K \tag{2.16}$$

eşitliği ile hesaplanır. Alt kitleler içerisinde en büyük ve en küçük değişim katsayıları

$$DK_{max} = max(DK_1, DK_2, \cdots, DK_K) \tag{2.17}$$

ve X sürekli rassal değişkeni için genel değişim katsayısı DK_{Genel} olmak üzere aşağıda tanımlanan değişim katsayısı oranı

$$DK_{oran} = \frac{DK_{Genel}}{DK_{max}} \tag{2.18}$$

özellik seçiminde kullanılabilir. İyi tanımlanmış bir sınıflamada, sınıf içi değişimin az, sınıflar arası değişim fazla olması gerekir. Genel değişim katsayısının büyük değer alması değişkenliğin fazla olduğunu gösterirken, sınıflara ait değişim katsayısının değerinin küçük olması sınıf içi değişkenliğin az olduğunu gösterir. Dolayısıyla yeni tanımlanan kriter için, kriterin büyük değer alması ilgili özelliğin sınıflamada daha etkili olabileceği anlamına gelir.

2.2. Karesel Diskriminant Analizi

Diskriminant analizi, en az hata ile birimleri ait oldukları sınıflara atamak için ayırma fonksiyonlarının oluşturulması ve oluşturulan ayırma fonksiyonları yardımı ile daha sonradan gözlemlenen, sınıf üyeliği bilinmeyen birimlerin sınıflandırılmasını gerçekleştirmeyi amaçlayan çok değişkenli istatistiksel bir analizdir (Rencher, 2003). Diskriminant analizinde sınıflara ait varyans-kovaryans matrisleri eşitse doğrusal diskriminant analizi ile ayırma fonksiyonları oluşturulmaktadır. Eğer sınıflara ait varyans-kovaryans matrisleri farklı ise yani $\Sigma_i \neq \Sigma_j$ durumunda karesel diskriminant analizi ile ayırma fonksiyonları oluşturulmaktadır. Sınıflara ait ortalama vektörü μ_k ve varyans-kovaryans matrisi Σ_k ve $k = 1, \dots, K$ olmak üzere karesel diskriminant fonksiyonu

$$Q_k(\mathrm{x}) = -\frac{1}{2} log|\Sigma_k| - \frac{1}{2}(\mathrm{x} - \mu_k)^T \Sigma_k^{-1}(\mathrm{x} - \mu_k) + log\pi_k \tag{2.19}$$

eşitliği ile oluşturulmaktadır. Eşitlikte yer alan π_k gösterimi k. gruba ait önsel olasılığı gösterir. Sınıflara ait ortalama vektörü μ_k ve varyans-kovaryans matrisi Σ_k olmak üzere Bayes kuralına göre x_i gözlem vektörünün k. sınıfa ait olma sonsal olasılığı

$$P(G = k|\mathrm{X} = \mathrm{x}_i) = \tau_{ik} = \frac{\pi_k f_k(\mathrm{x})}{\sum_{g=1}^{K} \pi_g f_g(\mathrm{x})} \tag{2.20}$$

eşitliği ile belirlenir. Eşitlikte yer alan $f_k(.)$ fonksiyonu ortalama vektörü μ_k ve varyans-kovaryans matrisi Σ_k olan çok değişkenli normal dağılıma ait olasılık yoğunluk fonksiyonudur. Eşitlik (2.20)'de verilen sonsal olasılığa göre x_i gözlem vektörü en yüksek olasılık değerinin elde edildiği sınıfa atanır.

2.3. Karşılaştırma Kriterleri

Bu çalışmada özellik seçim yöntemlerinin karesel diskriminant analizindeki sınıflama performansları sınıflama doğruluğu ve entropi kriterlerine göre karşılaştırılacaktır. Sınıflama performansının ölçümünde en yaygın kullanılan kriterlerden biri olan sınıflama doğruluğu, f_{kk} gerçek sınıf üyeliği k olup sınıflama analizi sonucunda doğru olarak k sınıfına atanan birimlerin sayısını göstermek üzere

$$CA = \frac{\sum_{k=1}^{K} f_{kk}}{n} \tag{2.21}$$

eşitliği ile elde edilir. Eşitlikte yer alan n toplam birim sayısını göstermektedir. Sınıflama doğruluğunun 1'e yakın olması sınıflama performansının başarılı olduğunu gösterir.

Sınıflama performansının ölçümünde kullanılan bir diğer ölçüt entropi, sınıflama belirsizliğinin bir ölçüsüdür. Entropi ölçütü n birimin K sınıfa sınıflandırılmasında

$$En(\tau) = -\sum_{i=1}^{n}\sum_{k=1}^{K} \tau_{ik} \ln(\tau_{ik}) \tag{2.22}$$

eşitliği ile hesaplanır. Entropi ölçütünün değeri ne kadar küçük ise sınıflama o kadar başarılıdır.

3. UYGULAMA

Önerilen özellik seçim yönteminin, iyi bilinen diğer özellik seçim yöntemlerinin karesel diskriminant analizindeki sınıflama performanslarının karşılaştırılması birim sayısının değişken sayısından az olduğu nicel verilerden oluşan elma, şekerpancarı ve Chowdary veri setleri üzerinde gerçekleştirilecektir.

Elma veri seti, Dedeoğlu (2011) tarafından elma ağaçlarında oluşan çinko eksikliğini görünür yakın kızılötesi yöntemle belirlendiği çalışma için oluşturulmuştur. Üç farklı bahçeden alınan elma yapraklarının spektral yansıma ölçümlerinden oluşan veri seti 60 birim ve 701 özellik

içermektedir. Şekerpancarı veri seti Dedeoğlu vd. (2019) tarafından şekerpancarı bitkisinin yapraklarındaki azot içeriğini hiperspektral yansımalar ile belirlediği çalışma için oluşturulmuştur. Veri seti şekerpancarı bitkisinin yapraklarındaki azot düzeyi noksan, yeterli ve fazla olmak üzere üç sınıf, 72 birim ve 601 özellikten oluşmaktadır. Çalışmada kullanılan Chowdary vd. (2006) tarafından oluşturulan üçüncü veri seti, lenf nodenegatif meme tümörleri ve Dukes'B kolon tümörlerinden elde edilen dokulardan oluşur. Veri seti, 62 meme tümörü ve 42 kolon tümörü olmak üzere 104 birim ve 22283 özellikten oluşmaktadır.

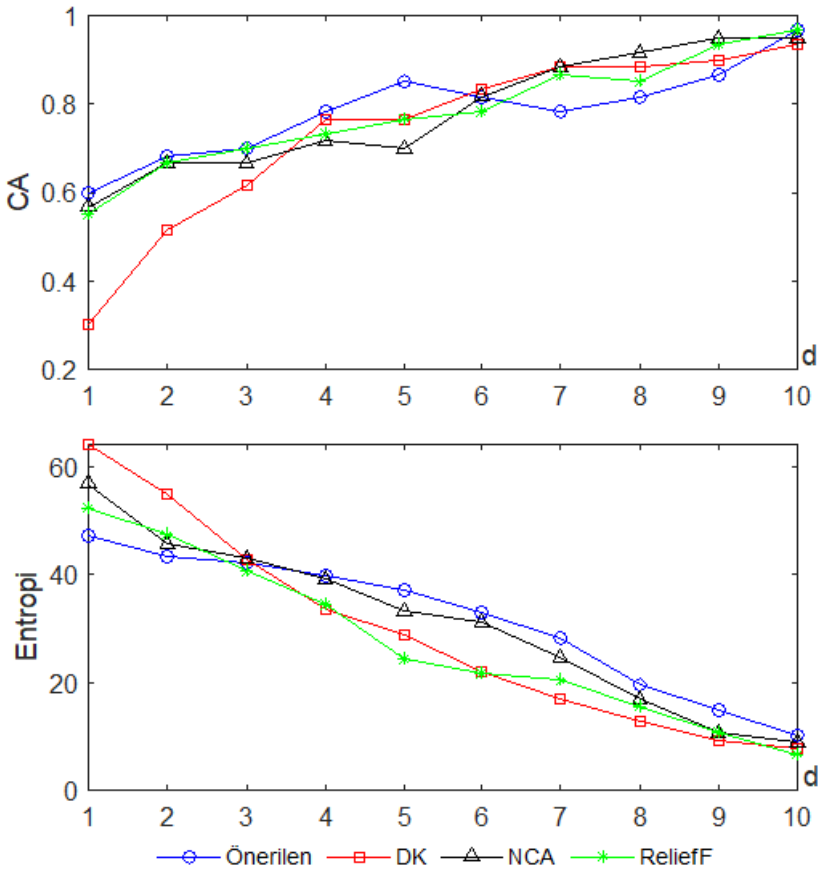
Çalışmanın uygulama aşamasında öncelikle incelenen özellik seçim yöntemlerinin her birine göre özelliklerinin önem sıralaması gerçekleştirilmiştir. Özelliklerin belirlenen önem sıralaması sonrasında en önemli bulunan özellik sayısı d olmak üzere her bir veri seti için ilk d özellik ile karesel diskriminant analizi uygulanmış ve analiz sonuçlarına göre sınıflama doğruluğu ve entropi değerleri hesaplanmıştır. Hesaplamalar MATLAB programında yazılan kodlar ile gerçekleştirilmiştir. Elma veri seti için elde edilen sonuçlar Tablo 1'de verilmiştir.

Tablo 1. Elma veri seti için karesel diskriminant analizinde kullanılan özellik sayısına göre özellik seçim yöntemlerinin sınıflama doğrulukları ve entropi değerleri

d	Sınıflama Doğruluğu						
	Önerilen	**DK**	**F**	**KW**	**NCA**	**ReliefF**	**Fisher**
1	0.600	0.300	0.550	0.550	0.567	0.550	0.550
2	0.683	0.517	0.633	0.617	0.667	0.667	0.633
3	0.700	0.617	0.633	0.617	0.667	0.700	0.683
4	0.783	0.767	0.717	0.667	0.717	0.733	0.783
5	0.850	0.767	0.667	0.683	0.700	0.767	0.817
6	0.817	0.833	0.683	0.717	0.817	0.783	0.800
7	0.783	0.883	0.800	0.767	0.883	0.867	0.833
8	0.817	0.883	0.917	0.850	0.917	0.850	0.917
9	0.867	0.900	0.900	0.867	0.950	0.933	0.883
10	0.967	0.933	0.933	0.917	0.950	0.967	0.917
d	**Entropi**						
	Önerilen	**DK**	**F**	**KW**	**NCA**	**ReliefF**	**Fisher**
1	47.270	64.129	52.369	52.704	56.807	52.369	52.369
2	43.333	54.943	46.748	50.588	45.655	47.380	46.748
3	42.151	42.651	43.729	45.571	43.106	40.551	43.515
4	39.831	33.581	38.267	42.689	39.077	34.431	37.324
5	37.136	28.763	33.077	38.343	33.151	24.427	34.206
6	32.991	22.031	29.911	31.279	31.056	21.644	28.927
7	28.251	16.914	25.706	23.487	24.689	20.391	24.610
8	19.503	12.716	16.025	17.092	16.886	15.560	19.905
9	14.800	9.363	13.819	15.327	10.586	10.802	15.609
10	10.249	7.640	10.576	10.861	8.869	6.652	9.640

Tablo 1 incelendiğinde sınıflama performansının ölçümünde kullanılan her iki kritere göre farklı başarı sıralamasının elde edildiği görülmektedir. Tablo 1'e göre boyut sayısı artıkça sınıflama performanslarının genel olarak iyileştiği ve yöntemler arasındaki farklılıkların azaldığı görülmüştür. Önerilen özellik seçim yöntemi, sınıflama doğruluğu bakımından boyut sayısı $d = 1, 2, 3, 4, 5, 10$ için seçili yöntemler arasında en başarılı yöntem iken entropi kriteri bakımından boyut sayısı $d = 1, 2$ için seçili yöntemler arasında en başarılı yöntem olmuştur.

Elma veri setinde, önerilen özellik seçim yönteminin değişim katsayısı, NCA ve ReliefF yöntemlerine karşı sınıflama performansının grafiksel gösterimi Şekil 1'de verilmiştir.



Şekil 1. Elma veri seti için karesel diskriminant analizinde kullanılan özellik sayısına göre seçili özellik seçim yöntemlerinin sınıflama performansları

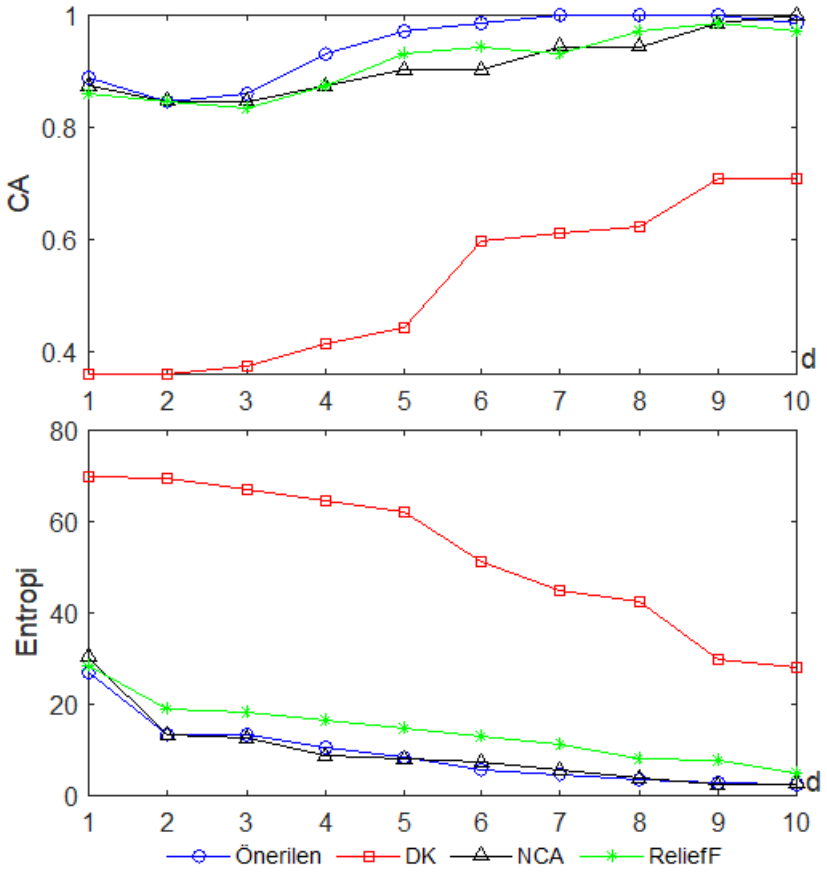
Şekil 1 incelendiğinde önerilen yönteminin seçili özellik seçim yöntemleri karşısında genel olarak başarılı bir performans gösterdiği görülmektedir. Elma veri seti için önerilen özellik seçim yöntemi özellikle boyut sayısının düşük olduğu durumlarda daha iyi sonuçlar vermiştir.

Şekerpancarı veri seti için karesel diskriminant analizinde kullanılan özellik sayısına göre özellik seçim yöntemlerinin sınıflama doğrulukları ve entropi değerleri Tablo 3.2'de verilmiştir.

Tablo 2. Şekerpancarı veri seti için karesel diskriminant analizinde kullanılan özellik sayısına göre özellik seçim yöntemlerinin sınıflama doğrulukları ve entropi değerleri

d	Sınıflama Doğruluğu						
	Önerilen	**DK**	**F**	**KW**	**NCA**	**ReliefF**	**Fisher**
1	0.889	0.361	0.875	0.861	0.875	0.861	0.861
2	0.847	0.361	0.847	0.861	0.847	0.847	0.875
3	0.861	0.375	0.875	0.861	0.847	0.833	0.861
4	0.931	0.417	0.875	0.903	0.875	0.875	0.903
5	0.972	0.444	0.903	0.903	0.903	0.931	0.903
6	0.986	0.597	0.903	0.917	0.903	0.944	0.931
7	1.000	0.611	0.931	0.944	0.944	0.931	0.944
8	1.000	0.625	0.958	0.972	0.944	0.972	0.958
9	1.000	0.708	0.931	0.972	0.986	0.986	0.944
10	0.986	0.708	0.972	0.972	1.000	0.972	0.972
d	**Entropi**						
	Önerilen	**DK**	**F**	**KW**	**NCA**	**ReliefF**	**Fisher**
1	26.894	69.810	26.072	24.608	30.306	28.297	24.608
2	13.500	69.355	14.446	18.156	13.315	19.089	19.369
3	13.376	67.139	14.985	16.808	12.605	18.368	16.808
4	10.655	64.436	14.214	15.008	8.770	16.534	15.008
5	8.315	62.060	12.700	14.801	8.029	14.659	14.130
6	5.794	51.186	12.130	12.216	7.357	12.903	11.502
7	4.746	45.013	10.653	8.845	5.498	11.216	8.204
8	3.442	42.329	8.796	6.518	3.935	8.100	6.485
9	2.885	30.018	6.672	6.116	2.512	7.637	5.405
10	2.407	28.206	4.756	5.023	2.651	4.949	3.857

Tablo 2 incelendiğinde sınıflama doğruluğu bakımından önerilen yöntem $d = 1, 4, 5, 6, 7, 8, 9$ için seçili yöntemler arasında en başarılı yöntem iken entropi kriteri bakımından boyut sayısı $d = 6, 7, 8, 9, 10$ için seçili yöntemler arasında en başarılı yöntem olmuştur. Önerilen özellik seçim yöntemi her iki kritere göre de klasik değişim katsayısı yöntemine göre şekerpancarı veri setinde belirgin bir şekilde daha iyi bir sınıflama performansı göstermiştir. Şekerpancarı veri setinde, önerilen özellik seçim yönteminin değişim katsayısı, NCA ve ReliefF yöntemlerine karşı sınıflama performansının grafiksel gösterimi Şekil 2'de verilmiştir.



Şekil 2. Şekerpancarı veri seti için karesel diskriminant analizinde kullanılan özellik sayısına göre seçili özellik seçim yöntemlerinin sınıflama performansları

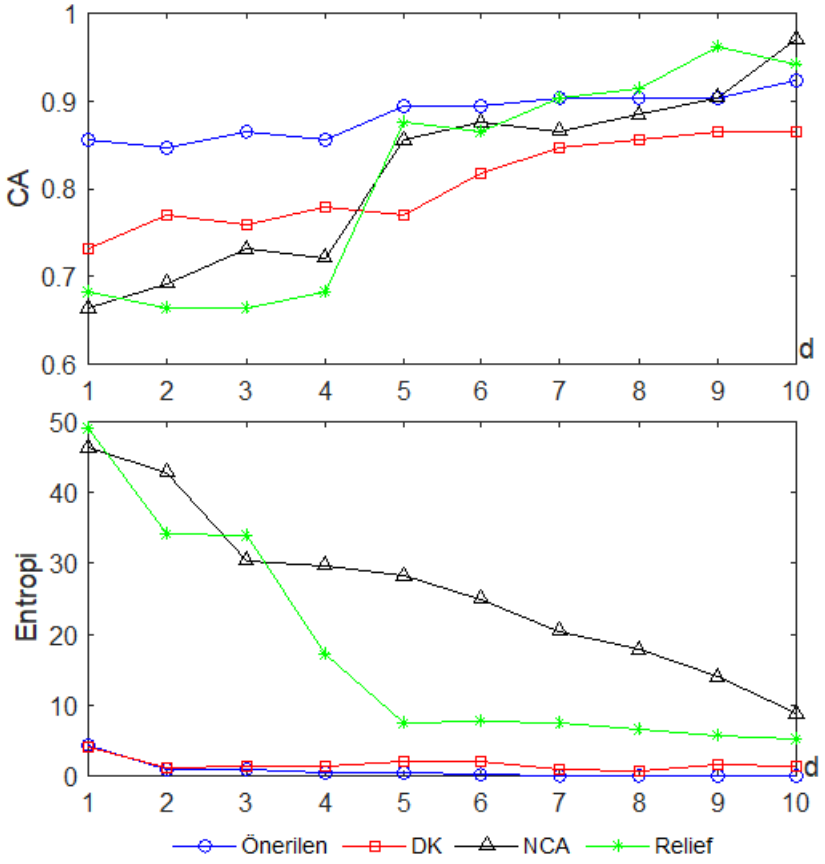
Şekil 2 incelendiğinde, değişik katsayısı yaklaşımının seçili özellik seçim yöntemleri arasında hem sınıflama doğruluğu hem de entropi kriteri bakımından en başarısız özellik seçim yöntemi olduğu görülmektedir. Önerilen özellik seçim yönteminin, sınıflama doğruluğu bakımından seçili özellik seçim yöntemleri arasındaki sınıflama performansı daha başarılı bulunmuştur. Önerilen yöntemin entropi kriterine göre sınıflama performansı NCA yöntemi ile birlikte diğer seçili özellik seçim yöntemlerinden daha başarılı olmuştur.

Chowdary veri seti için karesel diskriminant analizinde kullanılan özellik sayısına göre özellik seçim yöntemlerinin sınıflama doğrulukları ve entropi değerleri Tablo 3'de verilmiştir.

Tablo 3. Chowdary veri seti için karesel diskriminant analizinde kullanılan özellik sayısına göre özellik seçim yöntemlerinin sınıflama doğrulukları ve entropi değerleri

d	Sınıflama Doğruluğu						
	Önerilen	**DK**	**F**	**KW**	**NCA**	**Relief**	**Fisher**
1	0.856	0.731	0.865	0.962	0.663	0.683	0.808
2	0.846	0.769	0.933	0.962	0.692	0.663	0.952
3	0.865	0.760	0.923	0.962	0.731	0.663	0.962
4	0.856	0.779	0.942	0.962	0.721	0.683	0.962
5	0.894	0.769	0.962	0.971	0.856	0.875	0.952
6	0.894	0.817	0.990	0.971	0.875	0.865	0.952
7	0.904	0.846	0.990	0.971	0.865	0.904	0.971
8	0.904	0.856	0.990	0.971	0.885	0.913	0.981
9	0.904	0.865	0.990	0.971	0.904	0.962	0.990
10	0.923	0.865	0.990	0.971	0.971	0.942	0.990
d	Entropi						
	Önerilen	**DK**	**F**	**KW**	**NCA**	**Relief**	**Fisher**
1	4.412	4.021	34.010	4.703	46.338	49.091	21.359
2	0.878	1.256	12.533	1.060	42.843	34.210	8.170
3	0.982	1.436	9.104	0.166	30.415	33.983	1.134
4	0.433	1.386	3.490	0.464	29.736	17.120	0.970
5	0.518	2.111	2.122	0.143	28.186	7.530	0.524
6	0.140	2.120	0.680	0.174	24.958	7.796	1.456
7	0.003	0.982	0.124	0.395	20.446	7.507	3.306
8	0.000	0.706	0.173	0.556	17.917	6.604	2.378
9	0.000	1.556	0.003	0.054	13.978	5.759	1.301
10	0.000	1.345	0.000	0.047	8.874	5.161	0.000

Tablo 3 incelendiğinde önerilen özellik seçim yöntemi, sınıflama doğruluğu bakımından en başarılı yöntem olarak ortaya çıkmamasına karşın en başarılı sınıflama performansı gösteren diğer özellik seçim yöntemlerine yakın bir sınıflama performansına sahiptir. Bu yorumun doğruluğu önerilen özellik seçim yönteminin diğer karşılaştırma kriteri olan entropi değerleri incelendiğinde görülmektedir. Önerilen özellik seçim yöntemi, entropi kriteri bakımından boyut sayısı $d = 2, 4, 6, 7, 8, 9, 10$ için seçili yöntemler arasında en başarılı yöntem olmuştur. Sınıflama doğruluğu bakımından F, KW ve Fisher yaklaşımları en başarılı özellik seçim yöntemleri olmuştur. Sınıf bilgisi kullanılarak değişim katsayısına dayalı önerilen özellik seçim yöntemi bu veri setinde de genel olarak klasik değişim katsayısına göre daha başarılı sonuçlar vermiştir. Chowdary veri setinde, önerilen özellik seçim yönteminin değişim katsayısı, NCA ve Relief yöntemlerine karşı sınıflama performansının grafiksel gösterimi Şekil 3'de verilmiştir.



Şekil 3. Chowdary veri seti için karesel diskriminant analizinde kullanılan özellik sayısına göre seçili özellik seçim yöntemlerinin sınıflama performansları

Şekil 3 incelendiğinde önerilen yönteminin seçili özellik seçim yöntemleri karşısında genel olarak başarılı bir performans gösterdiği görülmektedir. Entropi kriteri bakımından seçili yöntemler arasında önerilen özellik seçim yöntemi ve klasik değişim katsayısı daha başarılı olmuştur.

4. SONUÇ

Bu çalışmada sınıf bilgisi kullanılarak değişim katsayısına dayalı yeni bir özellik seçim yöntemi önerilmiştir. Önerilen özellik seçim yönteminin etkinliği, yüksek boyutlu gerçek veri setleri kullanılarak iyi bilinen istatistik temelli filtreleme özellik seçim yöntemleri ile karesel diskriminant analizindeki sınıflama performansı bakımından karşılaştırılması ile değerlendirilmiştir. Özellik seçim yöntemlerinin sınıflama performansının ölçümünde sınıflama

doğrululuğu ve entropi kriterleri kullanılmıştır. Birim sayısının özellik sayısından az olduğu nicel verilerden oluşan üç gerçek veri seti ile gerçekleştirilen karşılaştırmalar sonucu, önerilen yöntemin sınıflama performansının genel olarak iyi olduğu görülmüştür. Çalışma sonuçları, önerilen yöntemin hesaplama kolaylığı göz önüne alındığında, özellik seçimi için güçlü bir alternatif olduğunu ortaya koymuştur.

Çalışmada özellik seçim yöntemlerinin sınıflama performanslarının veri setine ve ölçüm kriterine göre farklılık gösterdiği görülmüştür. Dolayısıyla özellik seçimi ile boyut indirgeme gerçekleştirilen çalışmalarda, veri setine uygun özellik seçim yönteminin belirlenmesi için birkaç özellik seçim yönteminin farklı karşılaştırma kriterlerine göre karşılaştırılmasının yararlı olacağı öngörülmüştür.

ETİK BEYAN

"Yeni bir özellik seçim yöntemi ve özellik seçim yöntemlerinin sınıflama performanslarının karşılaştırılması" başlıklı çalışmanın yazım sürecinde bilimsel, etik ve alıntı kurallarına uyulmuş; toplanan veriler üzerinde herhangi bir tahrifat yapılmamış ve bu çalışma herhangi başka bir akademik yayın ortamına değerlendirme için gönderilmemiştir.

KAYNAKÇA

Budak, H. (2018), Özellik seçim yöntemleri ve yeni bir yaklaşım, *Süleyman Demirel Üniversitesi Fen Bilimleri Enstitüsü Dergisi*, 22, 21-31.

Castro B.M. Lemes R.B. Cesar J. Hünemeier T. and Leonardi F. (2018), A model selection approach or multiple sequence segmentation and dimensionality reduction, *Journal of Multivariate Analysis*, 319-330.

Chowdary D, Lathrop J, Skelton J, Curtin K et al. (2006), Prognostic gene expression signatures can be measured in tissues collected in RNAlater preservative, 8(1), 31-39.

Dedeoğlu, M. (2011), Elma ve kiraz ağaçlarında çinko noksanlığının görünür yakın kızılötesi (VNIR) spektroradyometrik yöntemle belirlenebilirliğinin araştırılması, Selçuk

Üniversitesi, Fen Bilimleri Enstitüsü Toprak Bilimi ve Bitki Besleme Anabilim Dalı Yüksek Lisans Tezi, Konya.

Dedeoğlu, M., Başayiğit, L. ve Erişoğlu, M. (2019), Şeker pancarı yapraklarında azot durumunun spektral diskriminant analizi ile belirlenmesi, *Toprak Bilimi ve Bitki Besleme Dergisi*, 7(2), 128-138.

Guo C. and Wu D. (2018), Feature dimensionality reduction for video affect classification: A comparative study, 2018 First Asian Conference on Affective Computing and Intelligent Interaction (ACII Asia), *IEEE,* 1-6.

Kira. K. and Rendell L. A. (1992b), A practical approach to feature selection', In: D. Sleeman and P. Edwards (eds.): Machine Learning: Proceedings of International Conference (ICML'92), 249–256.

Kononenko, I. (1994). Estimating Attributes: Analysis and Extensions of RELIEF. Machine Learning: ECML-94, European Conference on Machine Learning, Secaucus, 6-8 April 1994, 171-182.

Kuhn, M. ve Johnson, K. (2013), Applied predictive modelling, New York: Springer.

Rencher, A. C. (2003), Methods of multivariate analysis, John Wiley & Sons.

Saeys, Y., Inza, I., ve Larranaga, P. (2007), A review of feature selection techniques in bioinformatics, *Bioinformatics*, 23(19), 2507-2517.

Singh G.D.A.A., Balamurugan S.A.A. and Leavline E. J. (2016), Literature review on feature selection methods for high-dimensional data, *International Journal of Computer Applications.* 8887. Foundation of Computer Science.

Yang W. Wang K. and Zuo W. (2012), Neighborhood Component Feature Selection for High-Dimensional Data, *JCP*, 7(1), 161-168.

Yıldız E. ve Sevim Y. (2016), Comparison of linear dimensionality reduction methods on classification methods, Electrical Electronics and Biomedical Engineering (ELECO), 2016 National Conference, *IEEE*, 161-164.