Journal of the Faculty of Engineering and Architecture of Gazi University 34:4 (2019) 1939-1955

Mühendislik Mimarlık Fakültesi Dergisi
Journal of The Faculty of Engineering and Architecture of Gazi University
Elektronik / Online ISSN : 1304 - 4915
Basılı / Printed ISSN : 1300 - 1884

A new approach based on regression analysis and mathematical programming to multi-group classification problems

Mustafa İsa Doğan[1], Abdullah Orman[2], Mediha Örkcü[3], H. Hasan Örkcü[4*]

[1]Department of Industrial Engineering, Düzce University, Düzce, 81620, Turkey
[2]Department of Technology, Ankara Yıldırım Beyazıt University, Ankara, 06760, Turkey
[3]Department of Mathematics, Gazi University, Ankara, 06500, Turkey
[4]Department of Statistics, Gazi University, Ankara, 06500, Turkey

Highlights:

- A new approach for the multi group classification problems
- Using the regression analysis for the obtaining the classification scores
- Using the mathematical programming for the classifying of the units

Keywords:

- Classification problem
- mathematical programming
- regression analysis
- two-stage approach.

Article Info:

Research Article
Received: 02.04.2018
Accepted: 21.12.2018

DOI:

10.17341/gazimmfd.571643

Correspondence:

Author: H. Hasan Örkcü
e-mail: hhorkcu@gazi.edu.tr
phone: +90 312 202 1462

Graphical/Tabular Abstract

Classification is the problem of identifying to which of a set of categories a new observation belongs, on the basis of a training set of data containing observations whose category membership is known. In this study, for solving multi-group classification problems, a new two-stage hybrid classification method based on regression analysis and mathematical programming has been developed. In the first step of the proposed method, the classification score of each unit is estimated with the help of the linear regression equation for each unit. In the second step, the classification of the units is performed by the mathematical programming model based on clustering analysis.

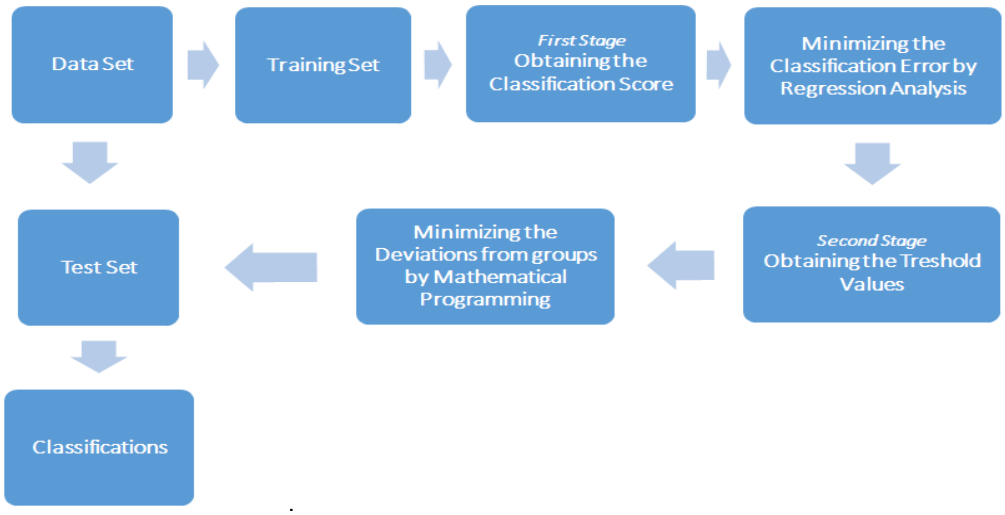


Figure A. Flow chart of proposed method

Purpose: In this study, for solving multi-group classification problems, a new method has been proposed based on regression analysis and mathematical programming classification method. The purpose of this study is handled the multi-group classification problem with the help of the superiority of regression analysis from the statistical theory and the flexibility of mathematical programming by a two-stage detailed examination idea.

Theory and Methods: The proposed method combines the strengths of regression analysis and mathematical programming method.

Results: From the 10 real data sets taken the well-known literature and simulation study results, it is observed that the proposed method outperforms the regression analysis, mathematical programming and artificial neural network based classification methods.

Conclusion: With the proposed method, it is possible to achieve high correct classification success in the multi-class classification problems.

Journal of the Faculty of Engineering and Architecture of Gazi University 34:4 (2019) 1939-1955

Mühendislik Mimarlık Fakültesi Dergisi
Journal of The Faculty of Engineering and Architecture of Gazi University

Elektronik / Online ISSN : 1304 - 4915
Basılı / Printed ISSN : 1300 - 1884

Çok gruplu sınıflandırma problemlerine regresyon analizi ve matematiksel programlama tabanlı yeni bir yaklaşım

Mustafa İsa Doğan[1], Abdullah Orman[2], Mediha Örkcü[3], H. Hasan Örkcü[4]*

[1]Düzce Üniversitesi, Konuralp Yerleşkesi, Mühendislik Fakültesi, Endüstri Mühendisliği Bölümü, Merkez, Düzce, 81620, Türkiye
[2]Yıldırım Beyazıt Üniversitesi, Meslek Yüksek Okulu, Yıldırım Beyazıt Mahallesi Ankara Bulvarı No:35 Çubuk, Ankara, Türkiye
[3]Gazi Üniversitesi, Fen Fakültesi, Matematik Bölümü, Teknikokullar, Ankara, 06500, Türkiye
[4]Gazi Üniversitesi, Fen Fakültesi, İstatistik Bölümü, Teknikokullar Ankara, 06500, Türkiye

Ö N E Ç I K A N L A R

- Çok gruplu sınıflandırma problemleri için yeni bir yaklaşım
- Sınıflandırma skorları için regresyon analizinin kullanılması
- Birimlerin sınıflandırılması için matematiksel programlamanın kullanılması

Makale Bilgileri
Araştırma Makalesi
Geliş: 02.04.2018
Kabul: 21.12.2018

DOI:
10.17341/gazimmfd.571643

Anahtar Kelimeler:
Sınıflandırma problem, matematiksel programlama, regresyon analizi, iki aşamalı yaklaşım

ÖZET

Bu çalışmada, çok gruplu sınıflandırma problemlerinin çözümü için regresyon analizi ve matematiksel programlamaya dayalı iki aşamalı yeni bir hibrit sınıflandırma yöntemi geliştirilmiştir. Önerilen yöntemin ilk aşamasında, her bir birimin sınıflandırma skoru her birim için oluşturulan doğrusal regresyon denklemi yardımıyla tahmin edilmektedir. İkinci aşamasında ise, birimlerin sınıflandırılması kümeleme analizi tabanlı matematiksel programlama modeli ile yapılmaktadır. Önerilen yöntem kendisini oluşturan regresyon analizi ve matematiksel programlama yöntemlerinin güçlü yanlarını kombine etmektedir. Literatürden alınan 10 gerçek veri seti ve simülasyon çalışması sonuçlarından, önerilen yöntemin regresyon analizi, matematiksel programlama ve yapay sinir ağı temelli sınıflandırma yöntemlerine göre daha iyi performans gösterdiği gözlemlenmiştir.

A new approach based on regression analysis and mathematical programming to multi-group classification problems

H I G H L I G H T S

- A new approach for the multi group classification problems
- Using the regression analysis for the obtaining the classification scores
- Using the mathematical programming for the classifying of the units

Article Info
Research Article
Received: 02.04.2018
Accepted: 21.12.2018

DOI:
10.17341/gazimmfd.571643

Keywords:
Classification problem, mathematical programming, regression analysis, two-stage approach

ABSTRACT

In this study, for solving multi-group classification problems, a new two-stage hybrid classification method based on regression analysis and mathematical programming has been developed. In the first step of the proposed method, the classification score of each unit is estimated with the help of the linear regression equation for each unit. In the second step, the classification of the units is performed by the mathematical programming model based on clustering analysis. The proposed method combines the strengths of regression analysis and mathematical programming method. From the 10 real data sets taken the well-known literature and simulation study results, it is observed that the proposed method outperforms the regression analysis, mathematical programming and artificial neural network based classification methods.

*Sorumlu Yazar/Corresponding Author: mustafaisadogan@duzce.edu.tr, aorman@ybu.edu.tr, medihaakcay@gazi.edu.tr, hhorkcu@gazi.edu.tr / Tel: +90 312 202 1462

1.GİRİŞ (INTRODUCTION)

Sınıflandırma problemi, en genel tanımı ile belirli bir sınıfa ait olduğu bilinen örneklerin kullanılarak, yeni örneklerin sınıflarının belli bir doğruluk ile belirlenmesidir [1]. Sınıflandırmanın başlıca uygulama alanları finansal yönetim ve ekonomi alanlarında; iflas risk tahminleri, kedi risk gruplarının belirlenmesi, yatırım araçlarında döviz, altın hisse senedi, tahvil vb. seçimi, ülkelerin kredi ve risk derecelendirmesi [2, 3], tıp alanında; gözlenen belirtilere göre hastanın uygun hasta grubuna atanması [4], mühendislik alanlarında; örüntü tanıma; insan özellikleri veya nesnelerin fiziksel özelliklerini tanıma ve sınıflandırılması [5, 6] ve bir sisteme gelen yabancı bir sinyalin belirlenmesi ya da uzaktan algılama ile arazi sınıflandırması olarak sıralanabilir [7-9].

Sınıflandırma problemlerinin çözümü için istatistiksel yöntemler, destek vektör makineleri, yapay sinir ağları, etkinlik analizi modelleri ve matematiksel programlama modelleri gibi birçok algoritma kullanılmaktadır. Fisher'in [10], iki grup için ortaya atmış olduğu istatistiksel doğrusal ayırma fonksiyonu, ayırma analizinde en eski yöntemdir. Fisher, çok değişkenli normal dağılım varsayımı altında iki ya da daha fazla gruptan gözlenmiş birimleri gruplardan birine sınıflandırmak için mevcut değişkenler üzerinden tanımlanacak doğrusal fonksiyonları önermiştir. Doğrusal ayırma fonksiyonunun katsayılar vektörü, gruplar arası farklılığı maksimum yapacak biçimde alınırlar. Ayırma fonksiyonu, $Y = \alpha_1 x_1 + \ldots + \alpha_k x_k = \underline{\alpha'} X$ olsun. İki gruplu doğrusal ayırma analizinde, k boyutlu α ayırma vektörü ve c sabiti belirlenir: $\underline{\alpha'} X < c$ ise ilgili birim 1. gruba, diğer durumda 2. gruba ait olarak sınıflandırılır. Bu doğrusal fonksiyon, farklı grupların kovaryans matrislerinin aynı olması durumunda uygundur. Aksi halde karesel fonksiyonların kullanımı analiz sonuçlarının sağlıklı olmasını sağlayacaktır [11]. İstatistiksel yöntemler kullanılarak elde edilen sonuçlar olasılıklarla ifade edilir fakat bu yöntemlerin kullanılabilmesi için gerekli olan varsayımların sağlanması çoğunlukla imkânsızdır [12]. Sınıflandırma problemlerinin incelenmesinde istatistiksel yöntemlere alternatif olarak çok sayıda matematiksel programlama yöntemi geliştirilmiştir. Matematiksel programlama yöntemlerini kullanmak için yığınla (kitle) ilgili herhangi bir varsayımın sağlanmasına gerek yoktur. Doğrusal programlama ile sınıflandırma problemlerinin incelenmesi ilk defa Fred ve Glover [13, 14] tarafından yapılmıştır. Fred ve Glover, iki gruplu sınıflandırma problemleri için, sapmalar toplamının minimizasyonuna dayanan bir model önermişlerdir. İki gruplu sınıflandırma probleminde X; G_1 ve G_2 gruplarından alınan n çaplı bir örneğin k tane değişken skorlarını gösteren $k \times n$'lik bir matris olsun. Değişken katsayıları $\alpha_1, \ldots, \alpha_k$ ile j. birime ait sınıflandırma skoru ise $S_j = \sum_{i=1}^{k} \alpha_i X_{ij}$, $(j = 1, \ldots, n)$ olarak tanımlanır. Bir birimin ait olduğu grubun tayin edilmesi o birimin sınıflandırma skorunun değerine bağlıdır. Fred ve Glover [13, 14] tarafından önerilen basit MSD (Minimization Sum of Deviations-Sapmalar Toplamının Minimizasyonu) sınıflandırma modeli,

$$\min \sum_{j \in G_1} S_j^+ + \sum_{j \in G_2} S_j^-$$

Kısıtlar:

$$\begin{aligned} &\textstyle\sum_{i=1}^{k} \alpha_i X_{ij} + S_j^+ \geq c, j \in G_1 \\ &\textstyle\sum_{i=1}^{k} \alpha_i X_{ij} - S_j^- \leq c, j \in G_2 \end{aligned} \quad (1)$$

Eş. 1 biçiminde tanımlanabilir. Burada $X_{ij} \geq 0$, $S_j^+ \geq 0$, $S_j^- \geq 0$ $(j = 1, \ldots, n)$, α_i $(i = 1, \ldots, k)$ ve c işaretçe serbest (pozitif veya negatif değerler alabilen) değişkenlerdir. Bu modelin çözülmesi ile α_i ve c değerleri elde edilerek, herhangi bir birimin sınıflandırma skoru elde edilebilir. Bir birimin sınıflandırma skoru (S_j), c'ye eşit ya da büyük ise G_1'e, diğer durumda G_2'ye sınıflandırılır. $S_j^+ \geq 0$ olması, G_1 grubundaki j. birimin yanlış sınıflandırıldığını; $S_j^- \geq 0$ olması, G_2 grubundaki j. birimin yanlış sınıflandırıldığını göstermektedir.

Matematiksel programlama yöntemleri değişkenler üzerinde herhangi bir dağılım varsayımına ihtiyaç duymamaktadırlar ve sınıflama amaçları için çok kullanışlıdırlar. Matematiksel programlama yöntemleri ile çoğunlukla iki gruplu sınıflandırma problemleri incelenmiştir. Lam vd. [15], kümeleme analizinde olduğu gibi aynı grup içindeki birimlerin farklı gruptaki birimlere göre daha benzer olması gerektiği fikrine dayalı olarak, doğrusal programlama tabanlı bir sınıflandırma yöntemi geliştirmiştir. Sueyoshi [16], DEA-DA adını verdikleri ilk aşamada çakışma durumunun tespit edildiği ikinci aşamada ise çakışma durumunda olan birimlerin sınıflandırıldığı iki aşamalı bir model önermiştir. Bu modelde veri zarflama analizi (VZA) modellerinden yararlanılmıştır. Sueyoshi [17], DEA-DA sınıflandırma modelini karma-tamsayılı formda ifade ederek yanlış sınıflandırılan toplam birim sayısını minimum yapmaya çalışan bir model önermiştir. Sueyoshi [18], karma-tamsayılı DEA-DA modelini orijinal DEA-DA ve diğer sınıflandırma algoritmaları ile bir simülasyon çalışması üzerinden karşılaştırmış ve veri yapısının özel bir durumu için DEA-DA modelini çok gruplu duruma genişletmiştir. Bal vd. [19], Lam vd. [15] modelinde ortalamadan sapmalar yerine medyan (ortanca) değerinden sapmalar toplamını minimize ederek özellikle çarpık dağılımdan gelen değişkenlere sahip problemlerde kullanılabilecek iki aşamalı LPMED modelini önermişlerdir.

Birden fazla sınıflandırma yönteminin güçlü yönlerinin bir araya getirilmesiyle oluşturulan hibrit sınıflandırma yöntemleri literatürde önemli bir yer tutmaktadır. Ramanan vd. [20], DAG-SVM adını verdikleri Dengesiz Karar Ağacı ve Destek Vektör Makinelerine dayalı hibrit bir yöntem önermişlerdir. Bal ve Örkcü [21], iki gruplu durum için MSD ve VZA yöntemlerinin bir kombinasyonu olan doğrusal programlamaya dayalı bir sınıflandırma modeli önermişlerdir. Önerilen model farklı dağılımlara göre yapılan simülasyon sonuçlarına göre MSD ve Fisher'in istatistiksel sınıflandırma fonksiyonundan daha yüksek

doğru sınıflandırma oranlarına sahiptir. Polat ve Güneş [22], çok gruplu sınıflandırma problemleri için C4.5 karar ağacı ve birine karşı hepsi yöntemlerinin bir kombinasyonu olan hibrit bir sınıflandırma yöntemi önermişlerdir. UCI veri tabanından alınan 3 farklı veri seti üzerinde önerilen yöntemin sınıflandırma başarısı doğru sınıflandırma oranı, duyarlılık ve 10-kat çapraz geçerlilik kriterlerine göre değerlendirilmiştir. Önerilen yöntem 3 veri setinde de C4.5 karar ağacı sınıflandırıcısından daha başarılıdır. Aci vd. [23], *k*-en yakın komşuluk, Bayes yöntemi ve genetik algoritmaya dayalı bir hibrit sınıflandırma yöntemi geliştirmişlerdir. UCI veri tabanından alınan 5 farklı veri seti üzerinde önerilen yöntemin sınıflandırma başarısı değerlendirilmiştir. Örkcü ve Bal [24] sınıflandırma problemlerinin çözümü için reel kodlu genetik algoritma ile eğitilmiş yapay sinir ağı modelini önermişlerdir. Khashei vd. [25], sınıflandırma problemleri için yapay sinir ağları ve regresyon analizine dayalı hibrit bir sinir ağı sınıflandırma yöntemi önermişlerdir. Bu yöntemde regresyon analizi yöntemi istatistiksel modellemedeki teorik üstünlüğü ile sinir ağlarındaki her bir ağırlık katsayısının büyüklüğünü belirlemek için kullanılmaktadır. UCI veri tabanından ve farklı kaynaklardan alınan veri setleri üzerinde önerilen yöntemin sınıflandırma başarısı 10-kat çapraz geçerlilik kriterine göre değerlendirilmiştir. Sonuçlar hibrit yöntemin doğrusal ayırma analizi, karesel ayırma analizi, klasik geri yayılım yapay sinir ağı, destek vektör makineleri ve *k*-en yakın komşuluk yöntemlerinden üstün olduğunu göstermektedir. Jabeen ve Baig [26], çok gruplu sınıflandırma problemleri için iki aşamalı bir öğrenme yöntemi önermişlerdir. İlk aşamada, sınıflandırıcılar kalan sınıflara karşı her sınıf için eğitilir. İkinci aşamada, sınıflandırıcılar, veri setinden herhangi bir sınıfı sınıflandırabilen tek bir kromozom olarak entegre edilmiş ve işlenmiştir. Chou vd. [27], çok gruplu sınıflandırma problemleri için bulanık mantık, genetik algoritma ve destek vektör makinaları yöntemlerinin bir kombinasyonu olan hibrit bir sınıflandırma yöntemi önermişlerdir. Önerilen yöntem Tayvan inşaat sektöründeki bir sınıflandırma problemine uygulanmış ve 10-kat çapraz geçerlilik kriterine göre kendisini oluşturan yöntemlerden daha başarılı olduğu ortaya konmuştur. Seera ve Lim [28], FMM-CART-RF adını verdikleri sınıflandırma ve regresyon ağaçları, bulanık min-max sinir ağları ve rasgele ağaç (random forest) yöntemlerine dayalı bir hibrit yöntem önermişlerdir. UCI veri tabanından alınan 3 farklı veri setine göre yapılan çalışmanın sonuçları önerilen hibrit yöntemin kendisini oluşturan sınıflandırma ve regresyon ağaçları ve bulanık min-max sinir ağları yöntemlerinden ve ayrıca literatürdeki bazı sınıflandırma yöntemlerinden 2-kat, 5-kat ve 10-kat çapraz geçerlilik kriterlerine göre daha yüksek sınıflandırma başarısına sahip olduğunu göstermiştir. Farid vd. [29], çok gruplu sınıflandırma problemleri için karar ağacı ve Bayes sınıflandırıcı yöntemlerine dayalı hibrit bir karar ağacı sınıflandırma yöntemi önermişlerdir. UCI veri tabanından alınan 10 farklı veri seti üzerinde önerilen yöntemin sınıflandırma başarısı 10 denemenin ortalamasına göre ve doğru sınıflandırma oranı, kesinlik, duyarlılık ve 10-kat çapraz geçerlilik kriterlerine göre değerlendirilmiştir. Sonuçlar, hibrit yöntemin karar ağacı ve Bayes sınıflandırıcı yöntemlerinden üstün olduğunu göstermektedir. Seera vd. [30], FAM-CART adını verdikleri bulanık ARTMAP yapay sinir ağları ve sınıflandırma ve regresyon ağaçları yöntemlerinin güçlü yönlerini birleştirerek hibrit bir sınıflandırma yöntemi önermişlerdir. UCI veri tabanından alınan medikal sınıflandırma veri setlerine ve Malezya'daki bir hastaneden alınan gerçek bir veri setine göre yapılan çalışmanın sonuçları önerilen hibrit yöntemin kendisini oluşturan sınıflandırma ve regresyon ağaçları ve ARTMAP yapay sinir ağları yöntemlerinden ve ayrıca yapay sinir ağları temelli bazı sınıflandırma yöntemlerinden kesinlik, duyarlılık ve 10-kat çapraz geçerlilik kriterlerine göre daha yüksek sınıflandırma başarısına sahip olduğunu göstermiştir. Kim ve Choi [31], HMC-LAD adını verdikleri ikili karar ağacına dayalı, verilerin mantıksal analizini kullanarak çalışan bir hiyerarşik çok gruplu sınıflandırma yöntemi önermişlerdir. Lee ve Lee [32], genetik algoritma ve ikili karar ağacı mimarisine dayalı hibrit bir destek vektör makineleri sınıflandırma yöntemi önermişlerdir. Bhardwaj vd. [33], çok gruplu sınıflandırma problemleri için genetik programlama ile optimize edilmiş sinir ağları yöntemini önermişlerdir. GONN adını verdikleri bu yöntemde her bir grup bir genetik programlama ağacı olarak ifade edilmektedir. UCI veri tabanından alınan 7 farklı veri seti üzerinde, farklı örnek bölümlemelerinde ve iki farklı sinir ağı mimarisinde önerilen yöntemin sınıflandırma başarısı klasik geri beslemeli yapay sinir ağı ve yapay sinir ağlarına dayalı sınıflandırma algoritmaları ile 10-kat çapraz geçerlilik kriterine göre değerlendirilmiştir. Önerilen yöntem klasik geri beslemeli yapay sinir ağı yöntemine göre oldukça başarılı sonuçlar vermiştir.

Bir sisteme gelen yabancı bir sinyalin belirlenmesi ya da uzaktan algılama ile arazi sınıflandırması gibi mühendislik problemleri için de çeşitli sınıflandırma yöntemleri geliştirilmiştir. Hedar vd. [34], bir sisteme izinsiz girişleri sınıflandırmak ve belirlemek için genetik algoritma, kaba set teorisi ve genetik programlamaya dayalı bir yöntem önermişlerdir. 25192 sinyal verisinin yanısıra UCI veri tabanından alınan sınıflandırma veri setlerine göre yapılan çalışmanın sonuçları doğru sınıflandırma oranları bakımından özellikle çok katmanlı sinir ağı ve destek vektör makinelerinden oldukça başarılı olduğunu ortaya koymaktadır. Lakshmi vd. [35], arazi kullanımı ve arazi sınıflandırması için karar ağacı ve istatistiksel doğrusal sınıflandırma yöntemine dayalı bir yöntem önermişlerdir. Zhang vd. [36], radyasyon sinyalini tanımlamak için, *k*-en yakın komşuluk, rasgele arama ve klasik geri yayılımlı yapay sinir sınıflandırma yöntemlerini kombine ederek hibrit bir sınıflandırma yöntemi önermişlerdir. Yöntemin doğru sınıflandırma başarısı karar ağacı, destek vektör makineleri, rasgele ağaç, *k*-en yakın komşuluk ve klasik geri yayılımlı yapay sinir ağı sınıflandırma yöntemleri ile 500 birimlik bir sinyal verisi üzerinden karşılaştırılmıştır. Önerilen hibrit yöntem sinyal tespitinde diğer sinir ağı sınıflandırma yöntemleri kadar yüksek sınıflandırma başarısına sahiptir.

İki gruplu durum için çok sayıda matematiksel programlama tabanlı yöntem geliştirilmiş, bu yöntemler çeşitli gerçek

dünya problemlerinde ve simülasyon deneylerinde karşılaştırılmıştır. Simülasyon çalışmalarında farklı dağılımlar ve gruplardaki birim sayılarının farklı büyüklükleri dikkate alınmıştır. Bu çalışmada ise iki gruplu duruma göre nispeten daha az önerinin yapıldığı çok gruplu sınıflandırma problemlerine literatürde önerilen matematiksel programlama yaklaşımları incelenmiş ve yeni bir yöntem önerisi yapılmıştır. Bu çalışmada önerilen yeni yöntem literaretürde önerilen Satapathy vd. [37] ve Lam ve Moy [38] modellerine dayananmaktadır. Satapathy vd. [37] sınıflandırma yaklaşımında her birimin sınıflandırma skoru doğrusal regresyon denklemi ile tahmin edilmekte fakat birimlerin sınıflara atanması için açık bir kural bulunmamaktadır. Bu çalışma, Satapathy vd. [37] sınıflandırma yaklaşımının bu eksikliğini Lam ve Moy [38] tarafından önerilen eşik değer sınaması yapan matematiksel programlama modeli ile gidermektedir. Lam ve Moy [38] tarafından önerilen çok gruplu sınıflandırma modelinde de birimlerin sınıflandırma skoru iki gruplu durumdaki MSD yaklaşımına benzer bir şekilde yapılmakta ve doğrusal programlama ile elde edilmektedir. Bu çalışmada önerilen model birimlerin sınıflandırma skorlarını bir istatistiksel modelleme yöntemi olan regresyon analizi ile elde etmesi bakımından Lam ve Moy [38] yaklaşımına katkı sunmakta, birimlerin sınıflandırmasını ise matematiksel programlama ile objektif bir şekilde elde etmesi bakımından da Satapathy vd. [37] yaklaşımına katkı sunmaktadır. Bu bakımdan bu çalışmada önerilen yöntem Satapathy vd. [37] ve Lam ve Moy [38] yöntemlerinin güçlü yanlarını birleştirmektedir.

Önerilen yöntemin ve bu çalışmada incelenen diğer çok gruplu sınıflandırma yöntemlerinin doğru sınıflandırma performansları literatürdeki gerçek veri setleri ve tasarlanan bir simülasyon çalışması ile karşılaştırılmıştır. Gerçek veri setleri için elde edilen sonuçlar literatürdeki hibrit yöntemlerin sonuçları ile de karşılaştırmalı olarak verilmiştir. Simülasyon çalışmasında önceki simülasyon çalışmalarına benzer olarak, verilerin geldiği dağılımların ve gruplardaki birim sayılarının farklı olduğu durumlar dikkate alınmıştır. Karşılaştırma kriteri olarak ise 10-kat çapraz geçerlilik kriteri kullanılmıştır.

Literatürde önerilen çok gruplu matematiksel programlama tabanlı sınıflandırma modelleri 2. Bölümde, önerilen sınıflandırma modeli 3. Bölümde, modellerin karşılaştırılması 4. Bölümde yer almaktadır. Son bölümde ise elde edilen sonuçlar yer almaktadır.

2 ÇOK GRUPLU SINIFLANDIRMA PROBLEMLERİNDE MATEMATİKSEL PROGRAMLAMA YAKLAŞIMLARI
(MATHEMATICAL PROGRAMMING APPROACHES IN THE MULTI-GROUP CLASSIFICATION PROBLEMS)

İki gruplu sınıflandırma problemi için önerilen çok sayıda matematiksel programlama yaklaşımı olmasına rağmen, çok az sayıda çok gruplu matematiksel programlama yaklaşımı vardır. İki gruplu durumdan çok gruplu duruma genişletmenin en kolay yolu tüm iki gruplu kombinasyonları kullanmaktır [14, 15]. Herhangi bir iki gruplu formülasyon bu çok gruplu ayırma yönteminde kullanılabilir. İki gruplu durum için bir karma tamsayılı sınıflandırma modeli, Eş. 2 ile verilebilir [39].

$$\min \sum_{j=1}^{n} y_j$$

Kısıtlar:

$$\begin{aligned} &\textstyle\sum_{i=1}^{k} \alpha_i X_{ij} - M y_j \le c - \varepsilon, j \in G_1 \\ &\textstyle\sum_{i=1}^{k} \alpha_i X_{ij} + M y_j \ge c + \varepsilon, j \in G_2 \end{aligned} \quad (2)$$

Burada, X_{ij}, j. birimin i. değişkenine ilişkin gözlem değeri, α_i, i. değişkenin ağırlığı (katsayısı), y_j, j. birimin yanlış sınıflandırılıp sınıflandırılmadığını gösteren iki değerli bir değişken, c ise iki grubu birbirinden ayıran ayırma eşik değeri olmak üzere serbest değerli bir değişkendir. ε ise küçük bir pozitif sayıdır ve herhangi bir birimin ayırma fonksiyonu üzerinde olmaması için modele dahil edilmektedir. h grup sayısı olmak üzere, $G_1, \ldots, G_h$ gruplarının tüm ikili çiftlerinin $h(h-1)/2$ tane ayırma fonksiyonu kurmak için kullanılması gerekmektedir.

Gehrlein [40] ikiden fazla grup için tasarlanmış özel bir karma tamsayılı programlama modeli önermiştir. Genel tek fonksiyonlu sınıflandırma modeli (GSFC),

$$\min \sum_{j=1}^{n} y_j$$

Kısıtlar:

$$\begin{aligned} &\alpha_0 + \textstyle\sum_{i=1}^{k} \alpha_i X_{ij} - M y_j \le U_r, j \in G_r, r = 1, \ldots, h \\ &\alpha_0 + \textstyle\sum_{i=1}^{k} \alpha_i X_{ij} + M y_j \ge L_r, j \in G_r, r = 1, \ldots, h \\ &U_r - L_r \ge e', r = 1, \ldots, h \\ &L_r - U_t + M J_{rt} \ge e, r = 1, \ldots, h, r \ne t \\ &L_t - U_r + M J_{tr} \ge e, r = 1, \ldots, h, r \ne t \\ &J_{rt} + J_{tr} = 1, r = 1, \ldots, h, r \ne t \end{aligned} \quad (3)$$

Eş. 3 olarak tanımlanmaktadır. Burada, h; grup sayısı, α_0; bir kayma sabiti (bir eşik değeri), U_r; G_r grubuna atanan aralığın son üst noktası ($r = 1, \ldots, h$), L_r; G_r grubuna atanan aralığın son alt noktası ($r = 1, \ldots, h$), e'; bir gruba atanan aralığın minimum genişliği, e; bitişik aralıklar arasındaki aralığın (oluşacak boşluğun) minimum büyüklüğü, M; pozitif büyük bir sabit, k; değişken sayısı ve n; toplam birim sayısı olarak tanımlanmaktadır. J_{rt} $(r \ne t)$ değişkeni G_r grubu G_t grubundan önce ise 1 değerini, değilse 0 değerini alan iki değerli bir değişkendir. y_j değişkeni, j. birimin yanlış sınıflandırılıp sınıflandırılmadığını gösteren iki değerli bir değişkendir. α_i $(i = 1, \ldots, k)$, L_r ve U_r $(r = 1, \ldots, h)$ değişkenleri ise serbest değerli değişkenlerdir.

Eş. 3 ile verilen model, her bir birim için bir doğrusal $\alpha_0 + \sum_{i=1}^{k} \alpha_i X_{ij}$ ayırma skoru tanımlamakta ve böyle bir ayırma skorunun $[L_r, U_r]$ aralığının içine düşüp düşmediğini kontrol

etmektedir. Ayırma skor değeri gruplar arasındaki boşluğa düşen bir birim yanlış sınıflandırılmış sayılmaktadır [40].

Gehrlein [40] ve aynı zamanlarda Choo ve Wedley [41] bir genel çok fonksiyonlu sınıflandırma modeli (GMFC) önermişlerdir.

$$\min \sum_{j=1}^{n} y_j$$

Kısıtlar:

$$\alpha_{r0} + \sum_{i=1}^{k} \alpha_{ri} X_{ij} - \alpha_{t0} - \sum_{i=1}^{k} \alpha_{ti} X_{ij} + M y_j \geq e,$$
$$j \in G_r, r = 1, \ldots, h; r \neq t$$
$$j = 1, \ldots, n \quad (4)$$

Bu modelde (Eş. 4), α_{ri}; G_r grubundaki X_{ij} değişkeninin ağırlığı ve α_{r0}: G_r grubu için kayma sabiti (G_r grubu için bir eşik değeri) olarak tanımlanmaktadır. Burada, α_{ri} ağırlık değişkenleri işaretçe serbest değişkenlerdir. y_j değişkeni, j. birimin yanlış sınıflandırılıp sınıflandırılmadığını gösteren iki değerli bir değişkendir. α_{ri} $(i = 1, \ldots, k; r = 1, \ldots, h)$ ise serbest değerli değişkenlerdir. GMFC modeli bir birimi en büyük ayırma skoruna sahip grup içine sınıflandırmaktadır. Modeldeki kısıt ile her bir grup için bir bireysel ayırma fonksiyonu oluşturulmaktadır. Eğer grup sayısı ve değişken sayısı çok büyürse, GMFC modeli GSFC modelinden çok daha fazla serbest değişken ve kısıtlamaya sahip olacaktır [42].

Lam ve Moy [38], Lam vd. [15] tarafından geliştirilen iki gruplu sınıflandırma modelini çok gruplu duruma genişletmişlerdir. Lam vd. [15]'nin iki gruplu sınıflandırma modelinde sınıflandırma işlemi bireysel sınıflandırma skorlarının kendi grup ortalama skorlarından sapmalarının minimize edilmesi temeline dayanmaktadır. MLM olarak isimlendirilen modelin çok gruplu duruma genişletme işlemi ise aşağıda verilmektedir. i. değişkene ilişkin ortalama $\bar{x}_i^r = \frac{1}{n_r} \sum_{j \in G_r} X_{ij}$, $(r = 1, \ldots, h)$ olarak tanımlanır. Burada n_r, G_r grubundaki birim sayısıdır. n ise $n = n_1 + \ldots + n_h$ olmak üzere toplam birim sayısını ifade etmektedir. $u = 1, \ldots, h-1$, $v = u+1, \ldots, h$ olmak üzere, her bir (u, v) çifti için model

$$\min \sum_{j \in G_u, j \in G_v}^{n} d_j$$

Kısıtlar:

$$\sum_{i=1}^{k} \alpha_i \left(X_{ij} - \bar{x}_i^u\right) + d_j \geq 0, j \in G_u$$
$$\sum_{i=1}^{k} \alpha_i \left(X_{ij} - \bar{x}_i^v\right) - d_j \leq 0, j \in G_v$$
$$\sum_{i=1}^{k} \alpha_i (\bar{x}_i^u - \bar{x}_i^v) \geq 1 \quad (5)$$

Eş. 5'de tanımlanmaktadır. Bu modelde, α_i $(i = 1, \ldots, k)$ serbest değerli değişkenler ve $j \in G_u$ ve $j \in G_v$ için $d_j \geq 0$'dır. $\sum_{i=1}^{k} \alpha_i \left(X_{ij} - \bar{x}_i^u\right) + d_j \geq 0$ ve $\sum_{i=1}^{k} \alpha_i \left(X_{ij} - \bar{x}_i^v\right) - d_j \leq 0$ kısıtları ile r. gruptaki birimlerin sınıflandırma skorlarını r. grubun ortalama sınıflandırma skoruna mümkün olduğunca yaklaştırmaya yarar $(r = u, v)$.

Eş. 5 yardımıyla her (u, v) grup çifti için elde edilen α_i yardımıyla G_u ve G_v 'deki bireylerin S_j sınıflandırma skor değerleri bulunur. Ardından da $u = 1, \ldots, h-1$, $v = u+1, \ldots, h$ olmak üzere grupları ayırmada yararlanılacak c_{uv} değerleri

$$\min \sum_{j \in G_u} \sum_{u=1}^{h-1} \sum_{v=u+1}^{h} d_{juv} + \sum_{j \in G_v} \sum_{u=1}^{h-1} \sum_{v=u+1}^{h} d_{juv}$$

Kısıtlar:

$$S_{juv} + d_{juv} \geq c_{uv}, u = 1, \ldots, h-1,$$
$$v = u+1, \ldots, h, j \in G_u$$
$$S_{juv} - d_{juv} \leq c_{uv}, u = 1, \ldots, h-1,$$
$$v = u+1, \ldots, h, j \in G_v \quad (6)$$

modelin çözülmesiyle elde edilir. Burada, tüm c_{uv} değişkenleri serbest değerlidir ve tüm d_{juv} değerleri ise negatif olmayan değişkenler olarak tanımlanmıştır. Eş. 6 ile verilen modelde tüm kesme (eşik) değerleri yani c_{uv} değerleri eş zamanlı olarak elde edilmektedir.

Gochet vd. [43] çok gruplu sınıflandırma problemi için LP^q yaklaşımını önerdiler. Yanlış sınıflandırmanın gruba uyumsuzluk olarak tanımlandığı bu modelde toplam uyumsuzluk miktarı minimize edilmeye çalışılmaktadır. j. birimin uyumsuzluğu β_{rt}^j, uyumluluğu ise γ_{rt}^j değişkenleri ile temsil edilirse, toplam uyumsuzluğun minimize yapılmaya çalışıldığı için LP^q yaklaşımı Eş. 7'de verilmektedir.

$$\min \sum \sum \sum \beta_{rt}^j$$

Kısıtlar:

$$\beta_{rt}^j + (\alpha^r - \alpha^t) X_{jr} - \gamma_{rt}^j = \varepsilon$$
$$\sum \sum \sum (\gamma_{rt}^j - \beta_{rt}^j) = q$$
$$\gamma_{rt}^j, \beta_{rt}^j \geq 0 \quad (7)$$

Burada, α^r ve α^t serbest değerli değişkenlerdir.

Sueyoshi [18], iki grup için verdiği karma-tamsayılı modelin çok gruplu duruma genişlemesini vermiştir. Çok gruplu durum için ise genel bir model $r = 1, \ldots, h$ için Eş. 8 ile verilmektedir.

$$\min \sum_{r=1}^{h} \sum_{j \in G_r} y_j$$

Kısıtlar:

$$\sum_{i=1}^{k}(\lambda_i^+ - \lambda_i^-)X_{ij} - c_r + My_j \geq 0\ ,\ r = 1,\ldots,h-1\ ,\ j \in G_r$$
$$\sum_{i=1}^{k}(\lambda_i^+ - \lambda_i^-)X_{ij} - c_r - My_j \leq -\varepsilon\ ,\ r = 1,\ldots,h-1\ ,\ j \in G_{r+1}$$
$$\sum_{i=1}^{k}(\lambda_i^+ + \lambda_i^-) = 1$$
$$\xi_i^+ \geq \lambda_i^+ \geq \varepsilon\xi_i^+\ ,\ i = 1,\ldots,k$$
$$\xi_i^- \geq \lambda_i^- \geq \varepsilon\xi_i^-\ ,\ i = 1,\ldots,k$$
$$\xi_i^+ + \xi_i^- \leq 1\ ,\ i = 1,\ldots,k$$
$$\sum_{i=1}^{k}(\xi_i^+ + \xi_i^-) = k \quad (8)$$

Bu modelde, λ_i^+ ve λ_i^- $(i = 1,\ldots,k)$ negatif olmayan değişkenler c_r $(r = 1,\ldots,h-1)$ serbest değerli değişkenler y_j $(j = 1,\ldots,n)$, ξ_i^+ ve ξ_i^- $(i = 1,\ldots,k)$ ise 0 ve 1 değerlerini alabilen ikili değişkenlerdir. $\lambda_i = \lambda_i^+ - \lambda_i^-$ olmak üzere, birimler aşağıda verilen kurala göre sınıflandırılmaktadır.

$\sum_{i=1}^{k}\lambda_i^* X_{im} \geq c_1^*$ ise G_1 grubuna
$c_{r-1}^* - \varepsilon \geq \sum_{i=1}^{k}\lambda_i^* X_{im} \leq c_r^*$ ise G_r $(r = 1,\ldots,h-1)$ grubuna
$c_{h-1}^* - \varepsilon \geq \sum_{i=1}^{k}\lambda_i^* X_{im}$ ise G_h grubuna

sınıflandırma yapılmaktadır. Eş. 8 ile $h-1$ tane farklı (c_1^*'dan c_{h-1}^*'a kadar) ayırma skoru elde edilmektedir. Ayırma skorları aynı λ_i^* $(i = 1,\ldots,k)$ ağırlık değerleri ile elde edilmektedir. Bu model çözüldüğünde optimal çözümün $c_1^* > c_2^* > \ldots > c_{h-1}^*$ olarak sağlanması gerekir. Çözüm bu durumu sağlamıyorsa $c_1 > c_2 + \varepsilon, c_2 > c_3 + \varepsilon, \ldots, c_{h-2} > c_{h-1} + \varepsilon$ kısıtlarının bu modele eklenmesi gerekmektedir. Sueyoshi tarafından verilen karma-tamsayılı yaklaşım çok gruplu sınıflandırmanın özel bir türünü çözebilir. Burada özel bir türle kastedilen veri yapısının sıralı halde olmasıdır. Eğer veri seti gruplara göre sıralı halde değilse mümkün olmayan çözümlerle veya düşük sınıflandırma oranları ile karşılaşılabilir.

Satapathy vd. [37], parçacık sürü optimizasyonu ve doğrusal regresyon analizi tabanlı bir sınıflandırma yaklaşımı önerdiler. Regresyon modelinin katsayıları her bir veri seti için ayrı ayrı en küçük kareler ve parçacık sürü optimizasyon teknikleri kullanılarak tahmin edilmektedir. Bu yaklaşımda birimlerin sınıflandırılması uzaklık ölçülerine göre yapılmaktadır.

Bal ve Örkcü [44], Sueyoshi'nin DEA-DA modeli [18] ve Gochet [43]'in çok gruplu sınıflandırma modelinin bir kombinasyonu olan, birçok gruplu sınıflandırma modeli önerdiler.

$$\min \sum_{r=1}^{h}\sum_{t\neq r}^{h}\sum_{j=1}^{n} n_{rt}^j$$

Kısıtlar:

$$\alpha_{r0} + \sum_{i=1}^{k}\alpha_{ri}X_{ij} - \alpha_{t0} - \sum_{i=1}^{k}\alpha_{ti}X_{ij} + n_{rt}^j - p_{rt}^j \geq \varepsilon\ ,\ j \in G_r\ ,\ r = 1,\ldots,h\ ;\ r \neq t\ j = 1,\ldots,n$$
$$\sum_{r=1}^{h}\sum_{i=0}^{k}\alpha_{ri} = 1\ ,\ r = 1,\ldots,h\ ;\ r \neq t$$
$$\alpha_{ri} \geq 0\ ,\ r = 1,\ldots,h\ ,\ i = 0,\ldots,k \quad (9)$$

Eş. 9 ile bir birim en büyük sınıflandırma skoruna sahip olduğu gruba atanır. ε değeri çok küçük pozitif bir sayıdır ve herhangi bir birimin ayırma skorunun ayırma düzlemi üzerinde olmasını engellemek için modele konulmuştur. Modelde $r.$ grupta yer alan birimlerin $r.$ grup için oluşturulan $\alpha_{r0} + \sum_{i=1}^{k}\alpha_{ri}X_{ij}$ ayırma fonksiyonunun, $t.$ grup için oluşturulan $\alpha_{t0} + \sum_{i=1}^{k}\alpha_{ti}X_{ij}$ ayırma fonksiyonundan ayırımı yolu ile gruplarına atanması istendiğinden $\alpha_{r0} + \sum_{i=1}^{k}\alpha_{ri}X_{ij} - \alpha_{t0} - \sum_{i=1}^{k}\alpha_{ti}X_{ij} \geq \varepsilon$ olması amaçlanmaktadır. Bu kısıta negatif n_{rt}^j ve pozitif p_{rt}^j sapma değişkenleri ilave edilerek ve istenmeyen n_{rt}^j sapma değişkenlerinin toplamı minimum yapılmaya çalışılarak birimlerin uygun gruplarına sınıflandırılması yapılmaktadır. İstenmeyen n_{rt}^j negatif sapma değişkenlerinin minimum yapılmasıyla $\sum_{i=1}^{k}\alpha_{ri}X_{ij} - \alpha_{t0} - \sum_{i=1}^{k}\alpha_{ti}X_{ij} \geq \varepsilon$ olması mümkün olduğunca sağlanacak ve grupların birbirinden ayırımı yapılabilecektir.

Xu ve Papageorgiou [45] hiper kutu (HB) sınıflandırıcısı adını verdikleri karma-tamsayılı matematiksel programlama modelini önerdiler. Bir hiper kutusu aslında çok boyutlu bir dikdörtgendir ve boyutlar, veri setindeki toplam özellik sayısına eşittir. Bu yöntem, her bir sınıf için mümkün olduğunca çok sayıda örnek içeren bir dizi hiper kutu inşa etmeyi amaçlamaktadır. Farklı sınıflara ait hiper kutular birbiriyle çakışmayacak şekilde sınırlandırılmıştır. Maskooki [46], hiper kutu sınıflandırıcısının değiştirilmiş bir versiyonunu önermiştir. Yang vd. [47] hiper kutu sınıflandırıcısını yeniden ele alarak, her iterasyonda sınıflandırıcı ağırlıkların güncellendiği ve hesaplama maliyetini azaltmak için veri setinde farklı bölümlemelerin yapıldığı geliştirilmiş bir hiper kutu sınıflandırma yöntemi önermişlerdir.

3. ÖNERİLEN SINIFLANDIRMA MODELİ (THE PROPOSED CLASSIFICATION MODEL)

Bu bölümde, Satapathy vd. [37] ve Lam ve Moy [38] modellerine dayalı olarak önerilen iki aşamalı sınıflandırma yaklaşımı yer almaktadır. Önerilen yaklaşımın ilk aşamasında her bir birimin sınıflandırma skoru Satapathy vd. [37]'e benzer bir şekilde her birim için oluşturulan doğrusal regresyon denklemi yardımıyla tahmin edilmektedir. İkinci aşamada ise Lam ve Moy [38] tarafından kullanılan sınıflandırma skorlarına dayalı eşik değer sınaması ile sınıflandırma yapılması sağlanmaktadır.

Regresyon analizi modelinin parametrelerinin tahmini bağımlı değişkenin gerçek gözlem değeri ile tahmin edilen değeri arasındaki fark olarak ifade edilen artık kareler toplamının en küçüklenmesi prensibine dayanmaktadır. Y bağımlı değişkeni, X açıklayıcı değişkenleri, β parametre vektörünü ve ε hata terimlerinin vektörünü göstermek üzere, $Y = X\beta + \varepsilon$ doğrusal regresyon model denkleminin tahmini

$\hat{Y} = X\hat{\beta} + \hat{\varepsilon}$ olmakta ve model parametrelerinin tahminleri de en küçük kareler tekniği ile yani $\hat{\varepsilon}$ artık kareler toplamının karesinin en küçüklenmesi ile yapılabilmektedir. $\sum_{i=1}^{n} \hat{\varepsilon}^2 = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_1 - \ldots - \beta_k X_k)^2$ artık kareler toplamının minimizasyonu Satapathy vd. [35]'e benzer şekilde parçacık sürü optimizasyonu ile yapılmıştır. Parçacık Sürü Optimizasyonu (PSO) kuşların veya balıkların sosyal öğrenmesinden esinlenen popüler bir sezgisel arama algoritmasıdır ve çok çeşitli alanlardaki optimizasyon problemleri için başarıyla uygulanmıştır [48, 49].

Sınıflandırma problemlerinde birimlerin skorlarını elde etmek amaçlı kullanılacak regresyon model denkleminde Y bağımlı değişkeni sadece grup numaralarının kodlandığı bir kategorik değişkendir yani birinci grup 1, ikinci grup 2, üçüncü grup 3 olarak kodlanmıştır. Satapathy vd. [37] artık kareler toplamının en küçüklenmesi işleminde parçacık sürüsü optimizasyonunu kullanmış fakat birimlerin gruplara atanması sürecinde uzaklık ölçülerine bağlı bir yöntem önermesine rağmen, birimlerin nasıl sınıflandırıldığı açık olarak belli edilmemiştir.

Lam ve Moy [38] tarafından önerilen ve Bal vd. [44] tarafından da hedef programlama çerçevesinde genişletilen model, çalışmamızda ikinci aşamada kullanılmaktadır. İlk aşamada ayırma fonksiyonunun ağırlıklarının ve bu sayede de birimlerin sınıflandırma skorlarının elde edilebilmesi için regresyon analizi kullanılmaktadır. İkinci aşamada ise regresyon analizi ile elde edilen sınıflandırma skorları Lam ve Moy [38]'un geliştirdiği modelin ikinci aşamasında yer alan doğrusal programlama modelinde kullanılarak, grupları ayırmak için eşik değerler elde edilmekte ve birimlerin sınıflandırılması yapılmaktadır.

Doğrusal regresyon modeli ile birimlerin sınıflandırma skorları elde edildikten sonra her (u, v) grup çifti için G_u ve G_v'deki birimlerin S_j sınıflandırma skor değerleri $\hat{\beta}$ regresyon tahminleri olmak üzere $S_j = X\hat{\beta}$ ile bulunur. Ardından da $u = 1, \ldots, h-1$ ve $v = u+1, \ldots, h$ olmak üzere grupları ayırmada yararlanılacak c_{uv} değerleri Eş. 10'un çözülmesi ile elde edilir.

$$\min \sum_{j \in G_u} \sum_{u=1}^{h-1} \sum_{v=u+1}^{h} d_{juv} + \sum_{j \in G_v} \sum_{u=1}^{h-1} \sum_{v=u+1}^{h} d_{juv}$$

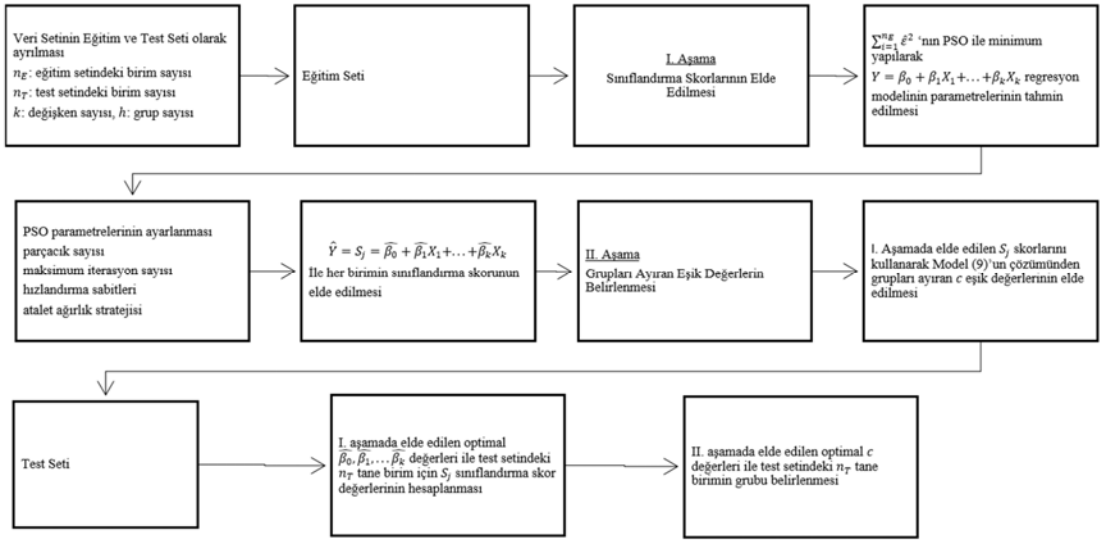
Kısıtlar:

$$\begin{aligned} &S_{juv} + d_{juv} \geq c_{uv}\,, u = 1, \ldots, h-1, \\ &v = u+1, \ldots, h\,, j \in G_u \\ &S_{juv} - d_{juv} \leq c_{uv}\,, u = 1, \ldots, h-1, \\ &v = u+1, \ldots, h\,, j \in G_v\; d_{juv} \geq 0 \end{aligned} \quad (10)$$

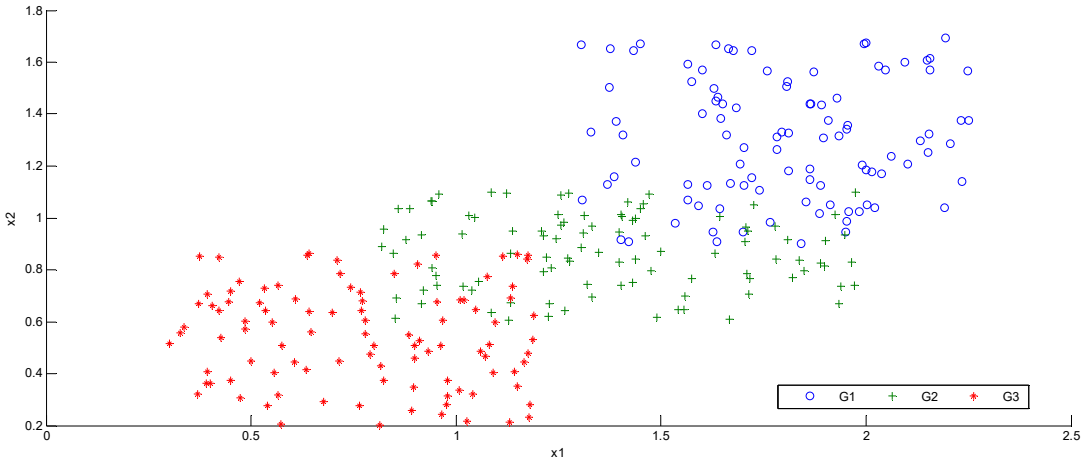
Burada, tüm c_{uv} değerleri serbest değerli değişkenler ve tüm d_{juv} değerleri ise negatif olmayan değişkenler olarak tanımlanmıştır. Bu modelde tüm kesme değerleri yani c_{uv} değerleri eş zamanlı olarak elde edilmektedir. Önerilen yaklaşımın akış diyagramı Şekil 1'de verilmektedir.

Önerilen yöntemin işleyişini göstermek için, birimlerin iki karakteristiğinin ölçüldüğü her grupta 100 birimin olduğu bir üç gruplu bir veri setini ele alalım. Veri setinin gruplara göre grafiği Şekil 2'de verilmektedir.

Birimlerin karakteristikleri x_1 ve x_2 değerleri olarak alınmış, y bağımlı değişkeni ise Satapathy vd. [37]'e benzer şekilde grup 1 birimleri için 1, grup 2 birimleri için 2 ve grup 3 birimleri için ise 3 olarak kodlanmıştır. $\sum_{i=1}^{300} (Y_i - \beta_0 - \beta_1 X_1 - \beta_2 X_2)^2$ artık kareler toplamı parçacık sürü optimizasyonu ile minimize edilerek, regresyon denklemi $\hat{y} = 4{,}0079 - 1{,}2143x_1 - 0{,}7018x_2$ olarak elde edilmiş ve bulunan regresyon denklemi yardımıyla her bir birimin sınıflandırma skoru elde edilmiştir. Örneğin ikinci grupta yer alan bir birimin x değerleri (1,2, 0,7) ise bu gruptaki bu birimin sınıflandırma skoru 2,059 olarak elde edilecektir. Üç gruptaki her birim için sınıflandırma skorları elde edildikten



Şekil 1. Önerilen yöntemin akış diyagramı (Flow chart of proposed method)



Şekil 2. Hipotetik veri setinin görsel gösterimi (Visual display of the hypothetical data set)

sonra grupları birbirinden ayıran ve birimlerin gruplarını tayin eden c eşik değerlerinin bulunması model (10)' un çözülmesiyle elde edilmiş ve eşik değerleri sırasıyla $c_1 = 1{,}5$ ve $c_2 = 2{,}7$ olarak elde edilmiştir. Bu küçük gösterim için, PSO işlemlerinde parçacık sayısı 20, maksimum iterasyon sayısı 100, hızlandırma sabitleri 2 ve atalet ağırlık stretejisi olarak ise 0,9 değerinden 0,1 değerine doğrusal bir şekilde azalan strateji kullanılmıştır [49].

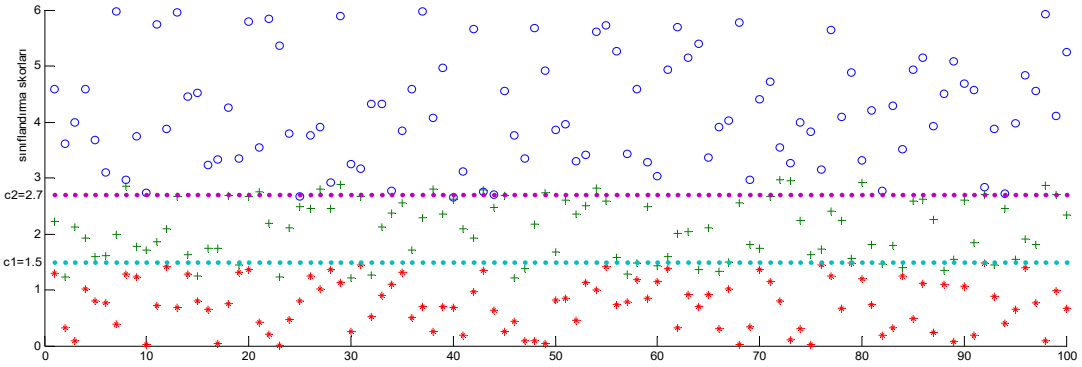
Üç gruptaki 100'er birimin sınıflandırma skorları ve sınıflandırma skorları sayesinde elde edilen grupları birbirinden ayıran eşik değerleri Şekil 3'de verilmektedir. Şekil 3'den, birimlerin elde edilen sınıflandırma skorları aracılığıyla doğrusal programlama yöntemi ile elde edilen sınıflandırma başarısının yüksek olduğu gözlenmektedir. Bu küçük örnekte birimlerin iki tane karakteristiği (x_1 ve x_2) alınmıştır. Ayrıca bu küçük görsel uygulamada veri seti bir bütün olarak alınmış yani verinin eğitim-doğrulama ya da k-tane alt gruba ayrıldığı performans ölçme durumları göz önüne alınmamıştır.

4. SINIFLANDIRMA MODELLERİNİN KARŞILAŞTIRILMASI
(COMPARISON OF THE CLASSIFICATION MODELS)

Sınıflandırma yöntemlerinin performanslarını karşılaştırmak için çapraz geçerlilik (cross validation) teknikleri kullanılmaktadır. Literatürde en çok kullanılan çapraz geçerlilik teknikleri; doğrulama örneği çapraz geçerlilik (test-holdout sample cross validation), bir birimi dışarıda tutma çapraz geçerlilik (leave-one-out cross validation), bazı birimleri dışarıda tutma çapraz geçerlilik (leave-some-out cross validation) ve $k-$ kat çapraz geçerlilik ($k-$fold cross validation) yaklaşımlarıdır [50, 51].

Doğrulama örneği çapraz geçerlilik yaklaşımında veri seti rasgele olarak iki sete ayrılır. Setlerden biri ayırma fonksiyonunun bulunuşunda kullanılır ve bu set eğitim veya geliştirme örneği (training sample, development sample) olarak isimlendirilmektedir. Diğer set ise doğrulama örneği (test-holdout sample) olarak isimlendirilir ve bu örnekler ayırma fonksiyonunun bulunuşunda kullanılmaz. Eğitim örneğinde ayırma fonksiyonu elde edildikten sonra yöntemin doğru sınıflandırma performansı doğrulama örneği ile sınanır. Bir birimi dışarıda tutma çapraz geçerlilik yaklaşımında bir birim veri setinden çıkartılır ve geriye kalan $n-1$ birim ile ayırma fonksiyonu elde edilir. Başlangıçta veri setinden çıkartılan birim $n-1$ birimden elde edilen ayırma fonksiyonu yardımıyla gruplardan birine sınıflandırılır. Bu işlem her bir birim için yani n defa tekrarlanır. Her birim için elde edilen doğru sınıflandırma sayılarının ortalaması ilgili yöntem için bir birimi dışarıda tutma çapraz geçerlilik doğru sınıflandırma performansı olarak alınır [52].

$k-$ kat çapraz geçerlilik yaklaşımında ise veri seti rasgele olarak k alt kümeye ayrılır. Bir birimi dışarıda bırakma yaklaşımına benzer olarak k alt kümeden bir tanesi veri setinden çıkartılır ve geriye kalan $k-1$ kümedeki birimlerle ayırma fonksiyonu elde edilir. Başlangıçta veri setinden çıkartılan küme içindeki birimler $k-1$ kümedeki birimlerden elde edilen ayırma fonksiyonu yardımıyla gruplara sınıflandırılırlar. Bu işlem her bir küme için yani k defa tekrarlanır [50, 53]. Her kümeden elde edilen doğru sınıflandırma sayılarının ortalaması ilgili yöntem için $k-$ kat çapraz geçerlilik doğru sınıflandırma performansı olarak alınır. Literatürdeki çalışmalarda k sayısı genellikle 10 alınmakta yani veri seti 10 alt kümeye ayrılmaktadır [28, 29, 33]. Bu çalışmada da veri seti rasgele olarak 10 alt kümeye ayrılarak sınıflandırma yöntemlerinin $10-$ kat çapraz geçerlilik doğru sınıflandırma performansları incelenmiştir. Regresyon doğrusu parametrelerinin tahmininde artık kareler toplamının minimizasyonu işleminde Satapathy vd. [37]'e benzer şekilde PSO'dan faydalanılmıştır. Gerçek veri setleri ve simülasyon çalışmasında, PSO işlemlerinde parçacık sayısı 100, maksimum iterasyon sayısı 1000, hızlandırma sabitleri $c_1 = c_2 = 2$ ve atalet ağırlık stratejisi olarak ise doğrusal azalan ağırlık stratejisi alınmıştır. Verinin büyüklüğüne ve problemin karmaşıklığına bağlı olarak parçacık sayısı 100 olarak belirlenmiştir. $c_1 = c_2 = 2$



Şekil 3. Birimlerin gruplara sınıflandırılması (Classification units to groups)

hızlandırma katsayıları ve doğrusal azalan atalet ağırlık stratejisi ise PSO uygulamalarında en çok kullanılan parametrelerdir [49].

4.1. Literatür Örnekleri (Literature Examples)

Fisher'in doğrusal ayırma fonksiyonu (FLDF), Gehrlein'in [40] çok fonksiyonlu sınıflandırma modeli (GMFC), Lam ve Moy [38]'un çok gruplu sınıflandırma modeli (MLM), Sueyoshi [18]'nin çok gruplu sınıflandırma modeli (DEA-DA), Gochet vd. [43] tarafından önerilen GCH modeli, Sapataty vd. [37] tarafından önerilen regresyona dayalı sınıflandırma modeli (SR), Bal ve Örkcü [44] tarafından önerilen hedef programlamaya dayalı sınıflandırma modeli (SGP), Yang vd. [47] tarafından geliştirilen hiper kutu sınıflandırma modeli (SHB) ve bu çalışmada önerilen çok gruplu sınıflandırma modeli yaklaşımlarının performansını değerlendirmek için 10 farklı veri seti dikkate alınmıştır. Ayrıca literatürde bu veri setlerini kullanan bazı hibrit yöntemler ile de karşılaştırmalı sonuçlara yer verilmiştir. 9 veri seti University of California-Irvine internet veri tabanından (UCI), 1 veri seti ise [54]'den alınmıştır ve kısaca izleyen şekilde açıklanan *IRIS*, *üzüm tanıma*, *cam tanımlama*, *ecoli*, *yeast protein tanımlama, mekanik analiz, dalga formu veri tabanı üreteci, Statlog mekik, Statlog görüntü segmentasyonu ve sahte twitter hesabı belirleme* gibi bazı gerçek hayat veri setlerini içermektedir [55].

1. FISHER'in IRIS Veri Seti: Bu veri setinde 3 farklı süs çiçeği türü yani 3 farklı grup vardır; *setosa*, *versicolor*, *virginica*. Veri seti üç grubun her birinde 50 çiçek olmak üzere toplam 150 çiçekten oluşmaktadır. Her bir çiçekten 4 farklı özellik gözlenmiştir; sepal uzunluğu, sepal genişliği, petal uzunluğu ve petal genişliği.

2. Üzüm Tanıma Veri Seti: Bu veri tabanında bulunan veriler, İtalya'nın aynı bölgesinde yetişen ancak üç farklı biçimde ekilen üzümlerin kimyasal bir analizinin sonuçlarıdır. Gruplarda yer alan üzümlerin sayısı sırasıyla 59, 71 ve 48'dir ve üzümler için 13 farklı kimyasal ölçüm yapılmıştır.

3. Cam Tanımlama Veri Seti: Cam veri seti, bir cam örnekleminin öz yapısını belirlemek için kullanılmaktadır. Cam türlerinin sınıflandırıldığı bu çalışma kriminolojik araştırma ile desteklenmiştir. Toplam 214 tane cam olmak üzere 6 farklı cam türü yani 6 farklı grup vardır. Gruplardaki cam sayıları sırasıyla 70, 17, 76, 13, 9 ve 29'dur ve camlardan 10 farklı kimyasal ölçüm yapılmıştır.

4. Ecoli-Protein Tanımlama Veri Seti: Ecoli veri seti, ecoli proteinlerinin hücresel konumlarının tahmin edilmesinde kullanılmaktadır. Toplam 327 tane ecoli proteini olmak üzere 5 farklı ecoli protein türü yani 5 farklı grup vardır (veri tabanında toplam 8 farklı grup olmasına rağmen 3 gruptaki birim sayısı çok az olduğundan 5 grup dikkate alınmıştır). Gruplardaki protein sayıları sırasıyla 143, 77, 52, 35 ve 20'dir ve proteinlerle ilgili 7 farklı kimyasal ölçüm yapılmıştır.

5. Yeast-Protein Tanımlama Veri Seti: Yeast veri seti de ecoli veri setine benzer olarak, yeast proteinlerinin hücresel konumlarının tahmin edilmesinde kullanılmaktadır. Toplam 1481 tane ecoli proteini olmak üzere 9 farklı ecoli protein türü yani 9 farklı grup vardır (veri tabanında toplam 10 farklı grup olmasına rağmen 1 gruptaki birim sayısı çok az olduğundan 9 grup dikkate alınmıştır). Gruplardaki protein sayıları sırasıyla 463, 429, 244, 163, 51, 44, 37, 30 ve 20'dir ve proteinlerle ilgili 8 farklı kimyasal ölçüm yapılmıştır.

6. Mekanik Analiz Veri Seti: Titreşim ölçümlerinden kaynaklı elektromekanik cihazlarda arızaların teşhisi ile ilgilidir. 209 cihazdan 8 farklı ölçüm alınmıştır ve arızalarla ilgili 6 grup içermektedir.

7. Dalga Formu Veri Tabanı Üreteci Veri Seti: Bu yapay veri seti, sınıflandırma amacı için simüle edilmiştir. Üç "temel" dalga üç grup oluşturmakta ve herbir örnek hepsi gürültü içeren 21 özellikten oluşmaktadır. Toplam 5000 örnek, üç gruba homojen olarak bölünmüştür.

8. Statlog Mekik Veri Seti: Bu veri NASA Shuttle veritabanından alınmış olup, uzay mekiğinde radyatörlerin konumlandırılması ile ilgilidir. Orijinalde 7 grup olmasına rağmen iki grupta çok az gözlem olmasından dolayı 5 grup kullanılmıştır. Toplamda 57700 örnekten 9 farklı özellik gözlemlenmiştir.

9. Statlog Görüntü Segmentasyonu Veri Seti: Yedi resimden (gruptan) olan bölümlerin çeşitli piksel ölçümleri ile değerlendirildiği bir veri setidir. Toplamda 2310 örnekten 19 farklı özellik ölçülmüştür.

10. Twitter Hesap Veri Seti: Twitterda açılan sahte hesapların belirlenmesi ile ilgili bir veri setidir. 48000 tane gerçek hesap 2000 tane de sahte olmak üzere 50000 örneğe (twitter hesabına) ait 10 farklı özellik ölçülmüştür (Verinin orjinali 95000 tane gerçek hesaptan 5000 tane ise sahte hesaptan oluşmaktadır, bu çalışmada 50000 tane twitter hesabı kullanılabilmiştir).

Tablo 1'de, sınıflandırma yöntemlerinin 100 tekrardan elde edilen ortalama $10-$ kat çapraz geçerlilik performansları ve ortalama doğru sınıflandırmalara ilişkin standart sapma değerleri yer almaktadır. Örneğin, Tablo 1'de, IRIS veri seti (set 1) için önerilen modele ait olan 0,959 ve 1,025 değerleri, sırasıyla 100 tekrardaki $10-$ kat çapraz geçerlilik doğru sınıflandırma ortalaması ve standart sapmasıdır. Önerilen yöntemin Yang vd. [47] hiper kutu yaklaşımı haricinde diğer yöntemlere göre 10 veri setinde de yüksek oranlarda doğru sınıflandırma başarısına sahip olduğu gözlenmektedir. Önerilen iki aşamalı yaklaşımın hiper kutu yaklaşımı ile beraber Cam Tanımlama verisi ve Dalga Formu Veri Tabanı Üreteci verisi için en yüksek doğru sınıflandırma oranlarına sahip olduğu, FLDF, GMFC, MLM, DEA-DA, GCH ve SR yöntemlerine göre ise daha başarılı sınıflandırma performansına sahip olduğu gözlenmektedir.

Ayrıca literatürde önerilen bazı hibrit yöntemlerin bu çalışmada ele alınan 10 veri seti için doğru sınıflandırma sonuçlarına da yer verilmektedir: Yang vd. [47] önerdikleri SHB yönteminin doğru sınıflandırma performansını literatürdeki bazı yöntemlerle iki senaryo üzerinden karşılaştırmışlardır. Senaryo 1, %70 eğitim seti %30 test seti veri bölümlemesine göre doğrulama örneği çapraz geçerlilik skorlarına, senaryo 2 ise bir birimi dışarıda tutma çapraz geçerlilik skorlarına göre oluşturulmuştur. IRIS veri seti için (set 1), Yang vd. [47]'nin elde ettiği sonuçlardan, SHB yöntemi senaryo 1'e göre %95,64, senaryo 2'ye göre %96 doğru sınıflandırma ortalamasına sahiptir. Bu çalışmada ise SHB yönteminin 100 tekrardaki $10-$ kat çapraz geçerlilik doğru sınıflandırma ortalaması %96,8 olarak elde edilmiştir. Yine Yang vd. [47]'nin elde ettiği sonuçlardan, Naive Bayes yöntemi için senaryo 1'e göre %96, senaryo 2'ye göre %95,33, Örkcü ve Bal [24] yöntemi için senaryo 1'e göre %86,93, senaryo 2'ye göre %88,67, Lojistik regresyon yöntemi için senaryo 1'e göre %95,56, senaryo 2'ye göre %98, Bagging yöntemi için senaryo 1'e göre %94,67, senaryo 2'ye göre %94, Adabost yöntemi için senaryo 1'e göre %94,36, senaryo 2'ye göre %97,33 ve klasik geri beslemeli yapay sinir ağları yöntemi için senaryo 1'e göre %95,11, senaryo 2'ye göre %95,33 doğru sınıflandırma ortalamaları elde edilmiştir.

IRIS veri seti için (set 1), Bhardwaj vd. [33]'nin elde ettiği sonuçlardan, 10- kat çapraz geçerlilik kriterine göre Ramanan vd. [20] tarafından önerilen DAG-SVM yöntemi %96,67, Jaben ve Baig [26] tarafından önerilen iki aşamalı hibrit yöntemi %96, Farid vd. [29] tarafından önerilen hibrit Naive Bayes yöntemi %98, Kim ve Choi [31] tarafından önerilen HMC-LAD yöntemi %96, Lee ve Lee [32] tarafından önerilen SVM yöntemi %97,27 ve Bhardwaj vd. [33] tarafından önerilen GONN hibrit yöntemi %97,44 doğru sınıflandırma ortalamasına sahiptirler.

Bu çalışmada önerilen iki aşamalı model, IRIS verisi için $10-$ kat çapraz geçerlilik kriterine göre %95,9 doğru sınıflandırma ortalamasına sahiptir.

Üzüm tanıma veri seti için (set 2), Bhardwaj vd. [33]'nin elde ettiği sonuçlardan, 10- kat çapraz geçerlilik kriterine göre Jaben ve Baig [26] tarafından önerilen iki aşamalı hibrit yöntemi %85, Farid vd. [29] tarafından önerilen hibrit Naive Bayes yöntemi %86,41 ve Bhardwaj vd. [33] tarafından

Tablo 1. Literatür veri setleri için yöntemlerin 10-kat çapraz geçerlilik doğru sınıflandırma oranları
(The correct classification ratios of 10-fold crossover validation of the methods for the literature data sets)

Yöntem	Set1	Set2	Set3	Set4	Set5	Set6	Set7	Set8	Set9	Set10
FLDF	0,855 (1,085)	0,831 (1,106)	0,701 (1,325)	0,776 (1,005)	0,710 (1,059)	0,718 (1,007)	0,777 (0,967)	0,779 (1,098)	0,796 (1,129)	0,807 (1,165)
GMFC	0,866 (1,051)	0,845 (1,082)	0,715 (1,209)	0,742 (1,089)	0,755 (1,106)	0,732 (1,116)	0,791 (1,130)	0,808 (1,107)	0,796 (1,109)	0,819 (1,187)
MLM	0,881 (1,141)	0,866 (1,028)	0,721 (1,341)	0,801 (1,222)	0,776 (1,111)	0,765 (1,119)	0,800 (1,120)	0,811 (1,119)	0,818 (1,001)	0,831 (1,099)
DEA-DA	0,893 (0,925)	0,866 (0,995)	0,741 (1,092)	0,799 (1,002)	0,781 (1,096)	0,799 (0,996)	0,802 (1,001)	0,845 (1,010)	0,844 (1,025)	0,855 (0,999)
GCH	0,899 (1,004)	0,871 (1,099)	0,778 (1,155)	0,802 (1,008)	0,811 (1,158)	0,822 (1,106)	0,811 (1,199)	0,861 (1,222)	0,859 (1,124)	0,874 (1,118)
SR	0,897 (1,325)	0,888 (1,199)	0,809 (1,389)	0,821 (1,209)	0,858 (1,199)	0,861 (1,201)	0,841 (1,009)	0,865 (1,108)	0,855 (1,007)	0,896 (1,192)
SGP	0,921 (1,201)	0,901 (1,155)	0,820 (1,146)	0,830 (1,072)	0,855 (1,011)	0,861 (1,002)	0,834 (1,145)	0,876 (1,008)	0,871 (1,109)	0,900 (1,138)
SHB	0,968 (1,011)	0,928 (0,978)	0,832 (1,109)	0,862 (0,992)	0,881 (1,009)	0,878 (1,112)	0,875 (1,113)	0,911 (1,008)	0,900 (1,009)	0,956 (1,111)
Önerilen Model	0,959 (1,025)	0,921 (1,001)	0,832 (1,129)	0,849 (1,029)	0,867 (0,996)	0,871 (1,008)	0,875 (1,023)	0,902 (0,979)	0,895 (0,991)	0,945 (0,895)

önerilen GONN hibrit yöntemi %91,66 doğru sınıflandırma ortalamasına sahiptirler.

Bu çalışmada önerilen iki aşamalı model, üzüm tanıma verisi için 10─ kat çapraz geçerlilik kriterine göre %92,1 doğru sınıflandırma ortalamasına sahiptir.

Cam tanımlama veri seti için (set 3), Yang vd. [47]'nin elde ettiği sonuçlara göre SHB yöntemi senaryo 1'e göre %71,09, senaryo 2'ye göre %66,36 doğru sınıflandırma ortalamasına sahiptir. Bu çalışmada ise SHB yönteminin 100 tekrardaki 10 ─ kat çapraz geçerlilik doğru sınıflandırma ortalaması %83,2 olarak elde edilmiştir. Yine Yang vd. [47]'nin elde ettiği sonuçlara göre, Naive Bayes yöntemi için senaryo 1'e göre %48,13, senaryo 2'ye göre %49,53, Örkcü ve Bal [24] yöntemi için senaryo 1'e göre %61,59, senaryo 2'ye göre %64,95, Lojistik regresyon yöntemi için senaryo 1'e göre %62,16, senaryo 2'ye göre %62,62, Bagging yöntemi için senaryo 1'e göre %68,69, senaryo 2'ye göre %72,90, Adabost yöntemi için senaryo 1'e göre %42,78, senaryo 2'ye göre %44,86 ve klasik geri beslemeli yapay sinir ağları yöntemi için senaryo 1'e göre %59,97, senaryo 2'ye göre %59,35 doğru sınıflandırma ortalamaları elde edilmiştir.

Cam tanımlama veri seti için (set 3), Bhardwaj vd. [33]'nin elde ettiği sonuçlardan, 10- kat çapraz geçerlilik kriterine göre Ramanan vd. [20] tarafından önerilen DAG-SVM yöntemi %65,54, Jaben ve Baig [26] tarafından önerilen iki aşamalı hibrit yöntemi %64, Farid vd. [29] tarafından önerilen hibrit Naive Bayes yöntemi %52,33, Lee ve Lee [32] tarafından önerilen SVM yöntemi %72,24 ve Bhardwaj vd. [33] tarafından önerilen GONN hibrit yöntemi %79,77 doğru sınıflandırma ortalamasına sahiptirler.

Bu çalışmada önerilen iki aşamalı model, cam tanımlama verisi için 10─ kat çapraz geçerlilik kriterine göre %83,2 doğru sınıflandırma ortalamasına sahiptir.

Ecoli protein tanımlama veri seti için (set 4), Bhardwaj vd. [33]'nin elde ettiği sonuçlardan, 10- kat çapraz geçerlilik kriterine göre Ramanan vd. [20] tarafından önerilen DAG-SVM yöntemi %81,99, Lee ve Lee [32] tarafından önerilen SVM yöntemi %84,52 ve Bhardwaj vd. [33] tarafından önerilen GONN hibrit yöntemi %83,22 doğru sınıflandırma ortalamasına sahiptirler.

Bu çalışmada önerilen iki aşamalı model, Ecoli protein tanımlama verisi için 10─ kat çapraz geçerlilik kriterine göre %84,90 doğru sınıflandırma ortalamasına sahiptir.

Yeast-protein tanımlama veri seti için (set 5), Bhardwaj vd. [33]'nin elde ettiği sonuçlardan, 10- kat çapraz geçerlilik kriterine göre, Jaben ve Baig [26] tarafından önerilen iki aşamalı hibrit yöntemi %64, Farid vd. [29] tarafından önerilen hibrit Naive Bayes yöntemi %71,59, Lee ve Lee [32] tarafından önerilen SVM yöntemi %59,82 ve Bhardwaj vd. [33] tarafından önerilen GONN hibrit yöntemi %75,88 doğru sınıflandırma ortalamasına sahiptirler.

Bu çalışmada önerilen iki aşamalı model, Yeast-protein tanımlama verisi için 10─ kat çapraz geçerlilik kriterine göre %86,7 doğru sınıflandırma ortalamasına sahiptir.

Mekanik Analiz veri seti için (set 6), Sangeetha ve Nalini [56]'nin elde ettiği sonuçlardan, doğrulama örneği çapraz geçerlilik kriterine göre, BayesNet yöntemi %84,98, Dagging yöntemi %88,35, Jrip yöntemi %87,58, J48 yöntemi %66,76 ve HyperPipes yöntemi %78,02 doğru sınıflandırma ortalamasına sahiptirler.

Bu çalışmada önerilen iki aşamalı model, mekanik analiz verisi için 10─ kat çapraz geçerlilik kriterine göre %87,1 doğru sınıflandırma ortalamasına sahiptir.

Dalga Formu Veri Tabanı Üreteci veri seti için (set 7), Xin-rong vd. [57]'nin elde ettiği sonuçlardan, doğrulama örneği çapraz geçerlilik kriterine göre, SVM-batch hibrit yöntemi %91,15, L1-KMSE-batch yöntemi %91,74, L1-KMSE-Increm yöntemi %91,09 ve SVM-Increm yöntemi %90,99 doğru sınıflandırma ortalamasına sahiptirler. Jia vd. [58]'nin elde ettiği sonuçlardan, 10─ kat çapraz geçerlilik kriterine göre, radyal tabanlı sinir ağları yöntemi %89, genetik algoritma ve radyal tabanlı sinir ağları yöntemlerine dayalı hibrit yöntem %87, genetik algoritma, radyal tabanlı sinir ağları ve en küçük karelere dayalı hibrit yöntem %97 doğru sınıflandırma ortalamasına sahiptirler.

Bu çalışmada önerilen iki aşamalı model, dalga formu veri tabanı üreteci verisi için 10─ kat çapraz geçerlilik kriterine göre %87,50 doğru sınıflandırma ortalamasına sahiptir.

Statlog Mekik veri seti için (set 8), Mendialdua vd. [59]'nin elde ettiği sonuçlardan, 5- kat çapraz geçerlilik kriterine göre, PPN K-NN hibrit yöntemi %99 doğru sınıflandırma ortalamasına sahiptir. Cimen vd. [60]'nin elde ettiği sonuçlardan, doğrulama örneği çapraz geçerlilik kriterine göre, *k*-ortalamalar tabanlı PCF yöntemi %96,50, ICF yöntemi %93,47 doğru sınıflandırma ortalamasına sahiptirler.

Bu çalışmada önerilen iki aşamalı model, Statlog mekik verisi için 10─ kat çapraz geçerlilik kriterine göre %90,20 doğru sınıflandırma ortalamasına sahiptir.

Statlog Görüntü Segmentasyonu veri seti için (set 9), Soybani [61]'nin elde ettiği sonuçlardan, 10- kat çapraz geçerlilik kriterine göre, AIRS2 hibrit yöntemi %88,24 ve FRA-AIRS2 yöntemi %89,65 doğru sınıflandırma ortalamasına sahiptirler.

Bu çalışmada önerilen iki aşamalı model, Statlog görüntü segmentasyonu verisi için 10─ kat çapraz geçerlilik kriterine göre %89,50 doğru sınıflandırma ortalamasına sahiptir.

Twitter Hesap veri seti için (set 10), Şimşek vd. [54]'nin elde ettiği sonuçlardan, 10- kat çapraz geçerlilik kriterine göre, Softmax-Softmax-Sigmoid yöntemi %97,72, Tanh-Tanh-

Sigmoid yöntemi %97,68, Rectifier- Rectifier-Sigmoid yöntemi %97,21, Rectifier-Rectifier-Tanh yöntemi %85,93, Tanh-Tanh-Rectifier yöntemi %95,94, Softmax-Softmax-Tanh yöntemi %95,24 ve Softmax-Softmax-Rectifier yöntemi %96,01 doğru sınıflandırma ortalamasına sahiptirler. Bu çalışmada önerilen iki aşamalı model, Twitter Hesap verisi için 10− kat çapraz geçerlilik kriterine göre %94,50 doğru sınıflandırma ortalamasına sahiptir. On veri seti beraber göz önüne alındığında, önerilen iki aşamalı hibrit yöntemin literatürde önerilen hibrit yöntemler ile rekabet edebilir durumda olduğu ve özellikle cam tanımlama, Ecoli, Yeast, Statlog mekik ve Statlog görüntü segmentasyonu veri setlerinde daha başarılı olduğu gözükmektedir.

4.2. Simülasyon Çalışması (Simulation Study)

Simülasyon çalışmasında çok gruplu durumlar olarak üç ve beş gruplu durumlar dikkate alınmıştır. Üç ve beş gruplu durumların her biri için üç değişkenli Düzgün (Uniform) ve Normal dağılımlardan veri üretilmiştir. Çalışmada kullanılan veri tipleri Tablo 2 ve Tablo 3'de özetlenmiştir.

Veri üretilirken veri yapısında varyanslara bağlı olarak değişkenlikler oluşturulmuştur. Örneğin $U[75{,}100]$ ve $U[25{,}85]$ dağılımlarının varyansları birbirinden önemli ölçüde farklıdır. Burada $U[75{,}100]$ ve $U[25{,}85]$ sırasıyla 75 ve 100 arasında ve 25 ve 85 arasında eşit olasılıklı olarak yani Düzgün olarak dağılan değerleri simgelemektedirler. $U[25{,}85]$ dağılımından alınan örneklerin değerleri $U[75{,}100]$ dağılımından alınan örneklere göre daha fazla değişkenliğe yani varyansa sahiptirler.

Aynı şekilde çok değişkenli normal dağılım durumunda da Σ varyans-kovaryans matrisindeki varyans ve kovaryansların gruplarda farklı değerler alması sağlanmıştır. Ayrıca gruplardaki birim sayıları için de iki farklı durum dikkate alınmıştır. n^A durumu tüm gruplardaki birim sayılarının eşit olduğu durumu göstermektedir ve bu durumda her gruptan 50'şer birim olacak şekilde veri üretimi sağlanmıştır. n^B durumu ise gruplardaki birim sayıların farklı olduğu durumu göstermektedir. Üç gruplu durumda $n_1 = 60$, $n_2 = 80$ ve $n_3 = 40$ olarak belirlenmiştir. Beş gruplu durumda ise $n_1 = 20$, $n_2 = 60$, $n_3 = 100$, $n_4 = 40$ ve $n_5 = 80$ olarak belirlenmiştir.

Simülasyon çalışmasında hem iki gruplu sınıflandırma problemleri için önerilen çalışmalardan [15, 19] ve hem de çok gruplu sınıflandırma problemleri için yapılan çalışmalardan faydalanılmıştır [38, 44]. Tablo 4'de, sınıflandırma yöntemlerinin 100 tekrardan elde edilen ortalama sınıflandırma ortalaması ve doğru sınıflandırma sayılarına ilişkin standart sapma değerleri yer almaktadır. Örneğin, Tablo 4'de, düzgün dağılım durumunda ve örnek çaplarının eşit olduğu durumda, üç gruplu durum için FLDF modeline ait olan 0,832 ve 1,234 değerleri, sırasıyla 100 tekrardaki FLDF modelinin doğru sınıflandırma ortalaması

Tablo 2. Üç gruplu durum için simülasyon çalışmasında kullanılan veri tipleri
(Data types used in the simulation study for the three group situation)

	Düzgün Dağılım	Normal Dağılım
Grup 1	$\{U[75{,}100], U[75{,}100], U[75{,}100]\}$	$\mu_1 = [75\ 75\ 75], \Sigma = \begin{bmatrix} 10 & 1 & 10 \\ 1 & 30 & 5 \\ 10 & 5 & 20 \end{bmatrix}$
Grup 2	$\{U[25{,}85], U[25{,}85], U[25{,}85]\}$	$\mu_2 = [50\ 50\ 50], \Sigma = \begin{bmatrix} 30 & 5 & 1 \\ 5 & 20 & 10 \\ 1 & 10 & 10 \end{bmatrix}$
Grup 3	$\{U[0{,}35], U[0{,}35], U[0{,}35]\}$	$\mu_3 = [40\ 40\ 40], \Sigma = \begin{bmatrix} 1 & 3 & 3 \\ 3 & 1 & 3 \\ 3 & 3 & 1 \end{bmatrix}$

Tablo 3. Beş gruplu durum için simülasyon çalışmasında kullanılan veri tipleri
(Data types used in the simulation study for the five group situation)

	Düzgün Dağılım	Normal Dağılım
Grup 1	$\{\mathrm{U}[100{,}150], \mathrm{U}[100{,}150], \mathrm{U}[100{,}150]\}$	$\mu_1 = [120\ 120\ 120], \Sigma = \begin{bmatrix} 1 & 3 & 3 \\ 3 & 1 & 3 \\ 3 & 3 & 1 \end{bmatrix}$
Grup 2	$\{\mathrm{U}[90{,}110], \mathrm{U}[90{,}110], \mathrm{U}[90{,}110]\}$	$\mu_2 = [100\ 100\ 100], \Sigma = \begin{bmatrix} 30 & 5 & 1 \\ 5 & 20 & 10 \\ 1 & 10 & 10 \end{bmatrix}$
Grup 3	$\{\mathrm{U}[75{,}100], \mathrm{U}[75{,}100], \mathrm{U}[75{,}100]\}$	$\mu_3 = [90\ 90\ 90], \Sigma = \begin{bmatrix} 10 & 1 & 10 \\ 1 & 30 & 5 \\ 10 & 5 & 20 \end{bmatrix}$
Grup 4	$\{\mathrm{U}[30{,}80], \mathrm{U}[30{,}80], \mathrm{U}[30{,}80]\}$	$\mu_4 = [70\ 70\ 70], \Sigma = \begin{bmatrix} 10 & 20 & 20 \\ 20 & 5 & 20 \\ 20 & 20 & 1 \end{bmatrix}$
Grup 5	$\{\mathrm{U}[0{,}40], \mathrm{U}[0{,}40], \mathrm{U}[0{,}40]\}$	$\mu_5 = [35\ 35\ 35], \Sigma = \begin{bmatrix} 1 & 30 & 10 \\ 30 & 20 & 1 \\ 10 & 1 & 1 \end{bmatrix}$

Tablo 4. Yöntemlerin 10-kat çapraz geçerlilik doğru sınıflandırma oranları
(10-fold cross validation correct classification ratios of the methods)

Yöntem	Örnek çapı durumu	Düzgün Dağılım		Normal Dağılım	
		Üç gruplu durum	Beş gruplu durum	Üç gruplu durum	Beş gruplu durum
FLDF	n^A	0,832 (1,234)*	0,811 (1,114)	0,849 (1,198)	0,831 (1,187)
	n^B	0,822 (1,245)	(0,802) (1,120)	0,833 (1,122)	0,827 (1,099)
GMFC	n^A	0,844 (1,101)	0,817 (0,993)	0,841 (1,003)	0,833 (1,006)
	n^B	0,827 (1,044)	0,802 (1,003)	0,836 (0,997)	0,830 (1,078)
MLM	n^A	0,861 (0,956)	0,842 (1,101)	0,862 (0,989)	0,855 (1,089)
	n^B	0,857 (0,988)	0,824 (1,095)	0,855 (1,056)	0,849 (1,045)
DEA-DA	n^A	0,865 (0,678)	0,854 (0,798)	0,867 (0,701)	0,858 (0,755)
	n^B	0,861 (0,699)	0,832 (0,806)	0,860 (0,755)	0,844 (0,789)
GCH	n^A	0,864 (1,065)	0,855 (1,007)	0,865 (1,032)	0,863 (1,001)
	n^B	0,852 (1,023)	0,844 (0,997)	0,855 (1,038)	0,850 (0,998)
SR	n^A	0,873 (2,432)	0,860 (2,654)	0,871 (2,006)	0,865 (2,023)
	n^B	0,868 (2,399)	0,845 (2,345)	0,870 (2,012)	0,851 (2,123)
SGP	n^A	0,911 (1,111)	0,883 (1,134)	0,913 (1,109)	0,899 (1,115)
	n^B	0,909 (1,006)	0,856 (1,078)	0,908 (0,999)	0,879 (1,009)
SHB	n^A	0,948 (1,006)	0,919 (1,109)	0,949 (0,999)	0,932 (1,008)
	n^B	0,941 (1,045)	0,900 (1,067)	0,946 (1,007)	0,916 (1,001)
Önerilen Model	n^A	0,936 (0,876)	0,911 (0,898)	0,939 (0,899)	0,926 (0,997)
	n^B	0,925 (0,903)	0,902 (0,911)	0,931 (0,942)	0,911 (0,956)

* 100 tekrardaki doğru sınıflandırma sayılarına ilişkin standart sapmalar

ve standart sapmasıdır. Önerilen modele ait 0,936 ve 0,876 değerleri de sırası ile 100 tekrardaki doğru sınıflandırma ortalaması ve standart sapmasıdır. SR yönteminin standart sapma değerleri her durum için diğer yöntemlere göre daha yüksek bulunmuştur. Buradan bu yöntemden elde edilen doğru sınıflandırma sayılarının değişkenliğinin diğer yöntemlere göre daha fazla olduğu söylenebilir. DEA-DA yöntemi için ise standart sapma değerleri her iki durumda da diğer yöntemlere göre daha küçük olarak elde edilmiştir. Bu durum, DEA-DA yönteminden elde edilen doğru sınıflandırma sayılarının daha tutarlı yani aynı dağılımdan alınan örnekler arasında çok büyük farklılıklar olmadığı şeklinde yorumlanabilir. Bu anlamda önerilen modelin doğru sınıflandırma sayılarına ilişkin standart sapma değerleri incelendiğinde, önerilen modelin de tutarlı sonuçlar verdiği söylenebilir. Bu bilgiler kullanılarak aşağıdaki hipotez testi yapılabilmektedir: H_0: Önerilen modelin doğru sınıflandırma ortalaması ile FLDF'nin doğru sınıflandırma ortalaması arasında fark yoktur. H_1: Önerilen modelin doğru sınıflandırma ortalaması FLDF'nin doğru sınıflandırma ortalamasından büyüktür. Aynı veri seti üzerinde iki farklı sınıflandırma yönteminin doğru sınıflandırma ortalamaları değerlendirildiğinden, yani sınıflandırma yöntemlerinin karşılaştırması bağımlı örnekler durumuna uygun olduğundan, ele alınan iki yöntemin doğru sınıflandırma ortalamaları arasında fark yoktur şeklindeki hipotezin testi bağımlı örneklemler t testi (paired samples t-test) ile yapılabilmektedir. İki yöntemin ortalamaları arasındaki fark $\bar{x}_d$ ve farklara ilişkin standart sapma s_d olmak üzere, n-1 serbestlik dereceli t istatistiğinin değeri $t_{n-1} = \bar{x}_d / s_d$ eşitliği ile hesaplanır. H_0 hipotezinin reddedilip edilmeyeceğine test istatiğinin değeri için hesaplanan p değerine göre karar verilir. p değeri analizden önce belirlenen α anlamlılık düzeyinden küçükse H_0 hipotezi reddedilecek ve ele alınan sınıflandırma yönteminin diğer yöntemden istatistiksel olarak daha yüksek sınıflandırma ortlamasına sahip olduğu söylenebilecektir. Bu çalışmada $\alpha = 0{,}01$ kabul edilmiştir ve 100 tekrar yapıldığı için $n = 100$'dür.

Tablo 5'de önerilen modelin doğru sınıflandırma ortalamasının, FLDF, GMFC, MLM, DEA-DA, GCH, SR,

Tablo 5. Önerilen modelin diğer (FLDF, GMFC, MLM, DEA-DA, GCH, SR, SGP ve SHB) yöntemlere karşı hipotez testi sonuçları (eşleştirme t değerleri)
(Hypothesis test results for the proposed method (paired t-values) against other (FLDF, GMFC, MLM, DEA-DA, GCH, SR, SGP and SHB)

Test	Örnek çapı durumu	Düzgün Dağılım		Normal Dağılım	
		Üç gruplu durum	Beş gruplu durum	Üç gruplu durum	Beş gruplu durum
Test 1 (FLDF)	n^A	12,01[a]	10,01[a]	6,33[a]	5,89[a]
	n^B	12,23[a]	10,22[a]	6,22[a]	6,01[a]
Test 2 (GMFC)	n^A	11,27[a]	10,41[a]	7,99[a]	7,16[a]
	n^B	11,49[a]	10,83[a]	8,12[a]	7,45[a]
Test 3 (MLM)	n^A	8,12[a]	9,01 [a]	6,77[a]	6,39[a]
	n^B	8,47[a]	9,22[a]	6,81[a]	6,31[a]
Test 4 (DEA-DA)	n^A	3,57[a]	3,41[a]	4,02[a]	3,92[a]
	n^B	3,55[a]	3,57[a]	4,18[a]	3,98[a]
Test 5 (GCH)	n^A	4,41 [a]	3,89 [a]	3,80[a]	3,77[a]
	n^B	4,39[a]	4,01[a]	3,89[a]	3,87[a]
Test 6 (SR)	n^A	4,04 [a]	3,08 [a]	3,52[a]	3,50[a]
	n^B	4,05[a]	3,09[a]	3,55[a]	3,47[a]
Test 7 (SGP)	n^A	2,69[a]	2,71[a]	2,75[a]	2,71[a]
	n^B	2,75[a]	2,83[a]	2,72[a]	2,69[a]
Test 8 (SHB)	n^A	-0,06	-0,05	-0,10	-0,15
	n^B	-0,11	-0,15	-0,16	-0,19

[a] H_0 Red ($p\ deg. < 0{,}01$)

SGP ve SHB modellerinin doğru sınıflandırma ortalamasından büyük olduğunu iddia eden hipotez testinin sonuçları yer almaktadır. Tablo 5'deki 12,01[a] değeri, düzgün dağılım durumunda ve örnek çaplarının eşit olduğu durumda, üç gruplu durum için, önerilen modelin doğru sınıflandırma ortalamasının FLDF yönteminin doğru sınıflandırma ortalamasından büyük olduğunun testindeki, t test istatistiğinin aldığı değeri gösterir. [a] ise bu hesaplanan 12,01'in p değeridir ve p değeri de 0,01' den küçük bir değerdir. Test istatistiğinin aldığı 12,01 değerine göre veya p değerine göre hipotezin sonucuna karar verilebilir. p değerinin 0,01' den küçük olması 0,01 anlamlılık düzeyinde H_0'ın reddedildiği sonucunu çıkarır. Sonuçta, istatistiksel olarak önerilen modelin doğru sınıflandırma ortalaması FLDF'nin doğru sınıflandırma ortalamasından büyük olduğu hipotezi kabul edilmiştir.

Tablo 5'deki elde edilen sonuçlar incelendiğinde önerilen modelin FLDF, GMFC, MLM, DEA-DA, GCH, SR ve SGP yaklaşımlarından her durumda üstün olduğu gözlenmektedir. Her ne kadar SHB yöntemi önerilen yöntemden daha iyi gözükse de istatistiksel olarak doğru sınıflandırma ortalamaları bakımından bir farklılık bulunamamıştır. Genel olarak, tüm sınıflandırma modelleri gruplardaki verilerin normal dağılımlı kitlelerden geldiği durumda düzgün dağılımlı duruma göre daha yüksek sınıflandırma performansı sergilemektedir. Beklendiği gibi, bu durum Fisher'in FLDF yönteminde çok bariz bir şekilde ortaya çıkmaktadır. Ayrıca tüm modeller tüm gruplarda eşit sayıda birim olduğu durumda gruplardaki birimlerin farklı olduğu duruma göre daha yüksek sınıflandırma başarısına sahiptirler.

5. SONUÇLAR (CONCLUSIONS)

Ayırma Analizi; tıp ve psikoloji bilim dallarında hastaların sınıflandırılması, biyoloji bilim dalında bitki türlerinin belirlenmesi, kimya bilim dalında maddelerin ayırımı ve mühendislik alanında bir sisteme gelen yabancı bir sinyalin belirlenmesi ya da uzaktan algılama ile arazi sınıflandırması gibi çok çeşitli alanlarda başarı ile kullanılan bir tekniktir. Bu çalışmada çok gruplu sınıflandırma problemlerinin çözümü için regresyon analizine ve matematiksel programlamaya dayalı iki aşamalı yeni bir hibrit sınıflandırma yöntemi geliştirilmiştir. Önerilen yöntem kendisini oluşturan regresyon analizi ve matematiksel programlama yöntemlerinin güçlü yanlarını kombine etmektedir. Literatürde sıklıkla kullanılan IRIS, üzüm tanıma, cam tanımlama, ecoli, yeast protein tanımlama, mekanik analiz, dalga formu veri tabanı üreteci, Statlog mekik, Statlog görüntü segmentasyonu ve twitter hesap veri setleri ile ve tasarlanan bir simülasyon çalışması ile, önerilen iki aşamalı yöntemin sınıflandırma performansı araştırılmıştır. Simülasyon çalışması, verilerin düzgün ve normal dağılımdan alındığı durumları ve gruplardan alınan örnek sayılarının eşit ve farklı olduğu durumları içermektedir. Literatürden alınan 10 gerçek veri seti ve simülasyon çalışması sonuçlarından, önerilen yöntemin, regresyon analizi, matematiksel programlama ve yapay sinir ağı temelli sınıflandırma yöntemlerine göre daha iyi performans gösterdiği gözlemlenmiştir.

TEŞEKKÜR (ACKNOWLEDGEMENT)

Twitterda açılan hesapların sahte ya da sahte olmadığı ile ilgili veri setini paylaştığı için Düzce Üniversitesi Mühendislik Fakültesi Bilgisayar Mühendisliği Bölümünden Dr. Mehmet ŞİMŞEK'e çok teşekür ederiz.

KAYNAKLAR (REFERENCES)

1. Silva, A.P.D., Optimization approaches to Supervised Classification, European Journal of Operational Research, 261, 772–788, 2017.

2. Mahmoudi, N, Duman, E., Detecting credit card fraud by Modified Fisher Discriminant Analysis, Expert Systems with Applications, 42, 2510–2516, 2015.

3. Yagi, T., Takahashi, M., Business cycle turning points based on DEA discriminant analysis, Applied Economics, 48 (44), 4251-4256, 2016.

4. Takcı, H., Centroid Sınıflayıcılar Yardımıyla Meme Kanseri Teşhisi, Journal of the Faculty of Engineering and Architecture of Gazi University, 31(2), 323-330, 2016.

5. Benmalek, E., Elmhamdi, J., Jilbab, A. Multiclass classification of Parkinson's disease using different classifiers and LLBFS feature selection algorithm. International Journal of Speech Technology, 20 (1), 179–184, 2017.

6. Yabanova, İ., Yumurtacı, M., Classification of dynamic egg weight using support vector machine, Journal of the Faculty of Engineering and Architecture of Gazi University, 33 (2), 393-402, 2018.

7. Jing, Z., Wang, G., Zhang, S., Qiu, C., Building Tianjin driving cycle based on linear discriminant analysis, Transportation Research Part D, 53, 78–87, 2017.

8. Wilson, S.R., Close, M.E., Abraham, B., Applying linear discriminant analysis to predict groundwater redox conditions conducive to denitrification, Journal of Hydrology, 556, 611–624, 2018.

9. Chen, L., Wang, E.Y., Feng, J.J., Wang, X.R., Li, X.L., Hazard prediction of coal and gas outburst based on Fisher discriminant analysis, Geomechanics and Engıneering, 13 (5), 861-879, 2017.

10. Fisher, R., The use of multiple measurements in taxonomic problems, Annals of Eugenics, 7 (2), 179-188, 1936.

11. Lachenburch P.A., Discriminant analysis, Hafner Press, New York, USA, 40-90, 1975.

12. Anderson, T.W., An introduction to multivariate analysis, Wiley, New York, USA, 10-25, 1984.

13. Fred, N., Glover, F., A Linear programming approach to the discriminant problem, Decision Sciences, 12, 68-74, 1981.

14. Fred, N., Glover, F., Simple but powerful goal programming formulations for the statistical discriminant problem, European Journal of Operational Research, 7, 44-60, 1981.

15. Lam, K.F., Choo, E.U., Moy, J.W., Minimizing deviations from the group mean: A new linear programming approach for the two-group classification problem, European Journal of Operational Research, 88, 358-367, 1996.

16. Sueyoshi, T., DEA-Discriminant analysis in the view of goal programming, European Journal of Operational Research, 115, 564-582, 1999.

17. Sueyoshi, T., Mixed integer programming approach of extended DEA-Discriminant analysis, European Journal of Operational Research, 152, 45-55, 2004.

18. Sueyoshi, T., DEA-Discriminant analysis: Methodological comparison among eight discriminant analysis approaches, European Journal of Operational Research, 169, 247-272, 2006.

19. Bal, H., Örkcü, H.H., Çelebioğlu S., "An alternative model to Fisher and linear programming approaches in two-group classification problem: minimizing deviations from the group median", G.U. Journal of Science, 19 (1), 49-55, 2006.

20. Ramanan, A., Suppharangsan, S.,Niranjan, M. Unbalanced decision trees for multi-class classification. In Industrial and information systems, 2007, 291-294, ICIIS 2007.

21. Bal, H., Örkcü, H.H., "Data envelopment analysis approach to two-group classification problems and an experimantal comparison with some classification models", Hacettepe Journal of Mathematics and Statistics, 36 (2), 169-180 , 2007.

22. Polat, K., Günes, S., A novel hybrid intelligent method based on C4.5 decision tree classifier and one-against-all approach for multi-class classification problems, Expert Systems with Applications, 36 (2), 1587–1592, 2009.

23. Aci, M., Inan, C., Avci, M., A hybrid classification method of k nearest neighbor, bayesian methods and genetic algorithm. Expert Systems with Applications, 37 (7), 5061–5067, 2010.

24. Örkcü, H. H., Bal, H., Comparing performances of backpropagation and genetic algorithms in the data classification. Expert Systems with Applications, 38 (4), 3703–3709, 2011.

25. Khashei, M., Hamadani, A. Z., Bijari, M., A novel hybrid classification model of artificial neural networks and multiple linear regression models. Expert Systems with Applications, 39 (3), 2606–2620, 2012.

26. Jabeen, H., Baig, A. R., Two-stage learning for multi-class classification using genetic programming. Neurocomputing, 116, 311–316, 2013.

27. Chou, J.S., Cheng, M.Y., Wu, Y.W., Improving classification accuracy of project dispute resolution using hybrid artificial intelligence and support vector machine models. Expert Systems with Applications, 40 (6), 2263–2274, 2013.

28. Seera, M., Lim, C.P., A hybrid intelligent system for medical data classification. Expert Systems with Applications, 41 (5), 2239–2249, 2014.

29. Farid, D. M., Zhang, L. , Rahman, C. M., Hossain, M. A., Strachan, R., Hy brid decision tree and naïve bayes classifiers for multi-class classification tasks. Expert Systems with Applications, 41 (4), 1937–1946, 2014.

30. Seera, M., Lim, C.P., Tan, S.C., Loo, C.K., A hybrid FAM-CART model and its applicatioon to medical data classification. Neural Computing and Applications, 26, 1799–1811, 2015.

31. Kim, H. H., Choi, J. Y., Hierarchical multi-class lad based on ova-binary tree using genetic algorithm. Expert Systems with Applications, 42 (21), 8134–8145, 2015.

32. Lee, Y., Lee, J., Binary tree optimization using genetic algorithm for multiclass support vector machine. Expert Systems with Applications, 42 (8), 3843–3851, 2015.

33. Bhardwaj, A, Tiwari, A., Bhardwaj, H., Bhardwaj, A., A genetically optimized neural network model for multi-

class classification. Expert Systems with Applications, 60, 211–221, 2016.
34. Hedar, A.R., Omer, M.A., Al-Sadek, A.F., Sewisy, A.A., Hybrid Evolutionary Algorithms for Data Classification in Intrusion Detection Systems, 2015 IEEE, SNPD 2015.
35. Lakshmi N. Kantakumar, Priti Neelamsetti, Multi-temporal land use classification using hybrid approach, The Egyptian Journal of Remote Sensing and Space Sciences, 18, 289–295, 2015.
36. Zhang, X., Luo, P., Hu, X., A Hybrid Method for Classification and Identification of Emitter Signals, The 4th International Conference on Systems and Informatics (ICSAI 2017).
37. Satapathy, S.C., Murthy, J.V.R., Prasad Reddy, P.V.G.D., Misra, B.B., Dash, P.K., Panda, G., Particle swarm optimized multiple regression linear model for data classification, Applied Soft Computing, 9, 470-476, 2009.
38. Lam, K.F., Moy, J.W., Improved linear programming formulations for the multi-group discriminant problem, Journal of Operational Research Society, 47, 1526-1529, 1996.
39. Pavur, R., Loucopoulos, C., Examining optimal criterion weights in mixed integer programming approaches to the multi group classification problem, Journal of Operational Researh Society, 46, 626-640, 1995.
40. Gehrlein, W.W., General mathematical programming formulations for the statistical classification problem, Operations Research Letters, 5, 299-304, 1986.
41. Choo, E.U., Wedley, W.C., Optimal criterion weights in repetitive multicriteria decision making, Journal of Operational Research Society, 36, 983-992, 1985.
42. Loucopoulos, C., Pavur, R., Computational characteristics of a new mathematical programming model for the three-group discriminant problem, Computers and Operations Research, 2, 179-191, 1997.
43. Gochet, W., Stam, A., Srinivisan, V., Chen, Shaoxiang, C., Multigroup discriminant analysis using linear programming, Operations Research, 45 (2), 213-225, 1997.
44. Bal, H., Örkcü, H.H., A new mathematical programming approach to multi-group classification problems, Computers and Operations Research, 38 (1), 105-111, 2011.
45. Xu, G., Papageorgiou, L. G., A mixed integer optimisation model for data classification. Computers & Industrial Engineering, 56, 1205–1215, 2009.
46. Maskooki, A., Improving the efficiency of a mixed integer linear programming based approach for multi-class classification problem. Computers & Industrial Engineering, 66, 383–388, 2013.
47. Yang, L., Liu, S., Tsoka, S., Papageorgiou, L.G., Sample re-weighting hyper box classifier for multi-class data classification, Computers & Industrial Engineering, 85, 44–56, 2015.
48. Eberhart, R., Kennedy, J. A new optimizer using particle swarm theory. In Micro Machine and Human Science, 1995. MHS'95., Proceedings of the Sixth International Symposium on, IEEE, 39-43, 1995.
49. Kennedy, J. Particle swarm optimization. In Encyclopedia of machine learning, Springer US, 760-766, 2011.
50. Izadi, B., Ranjbarian, B., Ketabi, S., Nassiri-Mofakham, F., Performance analysis of classification methods and alternative linear programming integrated with fuzzy delphi feature selection. International Journal of Information Technology and Computer Science (IJITCS), 5 (10), 9, 2013.
51. Zhang, Y., Yang, Y., Cross-validation for Selecting a Model Selection Procedure, Journal of Econometrics, 187, 95–112, 2015.
52. Vicente, T.F.Y., Hoai, M., Samaras, D., Leave-One-Out Kernel Optimization for Shadow Detection and Removal, IEEE Transactions on Pattern Analysis and Machine Intelligence, 40 (3), 682-695, 2018.
53. Kim B., Oertzen, T.V., Classifiers as a model-free group comparison test, Behav Res., 50, 416–426, 2018.
54. Şimşek, M., Yılmaz, O., Kahriman, A.H., Sabah, L., Sahte Twitter Hesaplarının Yapay Sinir Ağları ile Tespiti, Artificial Intelligence Studies, 1 (1), 19-22, 2018.
55. University of California-Irvine, www.ics.uci.edu/mlearn/MLRepository.html (2018).
56. Sangeetha, N., Nalini, C., "A Novel Approach: Various Classifications in Web Mining Using Diagnosis of Faults in Electro-Mechanical Devices From Vibration Measurements", International Journal of Mechanical Engineering and Technology (IJMET), 8 (10), 65-70, 2017.
57. Xin-rong, J., Cui-qin, H., Yi-bin, H., Da, L., "An Incremental Learning Method for L1- Regularized Kernel Machine in WSN", 3rd International Conference on Mechatronics and Industrial Informatics (ICMII 2015), 397-403.
58. Jia,W., Zhao,D., Shen, T., Su, C., Hu, C., Zhao, Y., "A New Optimized GA-RBF Neural Network Algorithm", Computational Intelligence and Neuroscience, 982045, 2014.
59. Mendialdua, I., Oses, N., Sierra, B., Lazkano, E., Positive Predictive Value based dynamic K-Nearest Neighbor, Advances in Knowledge-Based and Intelligent Information and Engineering Systems, Editor: Grafia, M., 59-68, 2012.
60. Cimen, E., Ozturk, G., Gerek, O.N., ICF: An algorithm for large scale classification with conic functions, SoftwareX, inpress, 77, 187-194, 2018.
61. Saybani, M. R., A Hybrid Approach For Artificial Immune Recognition System, Doctoral Thesis, Faculty of Computer Science and Informatıon Technology, Universıty of Malaya, Kuala Lumpur, 2016.