

Kohonen Öğrenme Kuralı Yardımıyla Centroid Değerleri Güncelleme

Updating Centroid Values Using Kohonen Learning Rule

Hidayet Takcı

Sivas Cumhuriyet Üniversitesi, Mühendislik Fakültesi, Bilgisayar Mühendisliği Bölümü, Sivas, Türkiye

ORCID: 0000-0002-4448-4284, htakci@cumhuriyet.edu.tr

Geliş Tarihi: 12 Ekim 2025 - Kabul Tarihi: 4 Kasım 2025

DOI: 10.55205/joctensa.3220241801338

ATIF: Takcı, H., (2025). Kohonen öğrenme kuralı yardımıyla centroid değerleri güncelleme. *Cihannüma Teknoloji Fen ve Mühendislik Bilimleri Dergisi*, 3 (2), 57-67.

ÖZ

Doğal dil işleme alanında öne çıkan görevlerden birisi metin sınıflandırmadır. Metin sınıflandırma için destek vektör makineleri, yapay sinir ağları, naive bayes gibi çok sayıda makine öğrenmesi algoritması kullanılmakta olup onlardan birisi de centroid tabanlı sınıflayıcıdır. Performansı naive bayes sınıflayıcı kadar iyi olan centroid sınıflayıcıların en büyük problemi kullanılan veriye aşırı bağımlılıktır. Bu problemin çözümü için en uygun çözümlerden birisi centroid değerlerinin güncellenmesidir. Centroid tabanlı sınıflayıcılar eğitim verilerinden centroid vektörlerinin elde edilmesi ve bu değerlerin statik olarak kullanılmasına dayalıdır. Statik değerler ile belli bir doğruluğun üzerine çıkmak mümkün olmadığı için centroid değerlerinin güncellenmesine ihtiyaç duyulur. Centroid güncelleme için bugüne kadar birçok yöntem geliştirilmiş olup bu çalışmada yeni bir yöntem önerilmektedir. Önerilen yöntem; Kohonen öğrenme kuralı yardımıyla centroid değerlerinin dinamik olarak güncellenmesine dayalıdır. Bu yöntemle göre centroid değerleri iteratif olarak güncellenmiştir. ECI CD ROM veri seti üzerinde yapılan deneylerde Kohonen öğrenme kuralının sınıflandırma performansını %0,56 oranında artırdığı gözlemlenmiştir. Bu durum Kohonen öğrenme kuralını centroid sınıflayıcıları geliştirmek için bir alternatif haline getirmektedir.

Anahtar Kelimeler: Centroid Sınıflayıcı, Model Misfit, Centroid Güncelleme, Kohonen Öğrenme Kuralı

ABSTRACT

Text classification is one of the most prominent tasks in natural language processing. Numerous machine learning algorithms, such as support vector machines, artificial neural networks, and naive Bayes, are used for text classification, and one of these is the centroid-based classifier. The biggest problem with centroid classifiers, which perform as well as naive Bayes classifiers, is their excessive dependence on the data used. One of the most suitable solutions to this problem is updating the centroid values. Centroid-based classifiers are based on obtaining centroid vectors from training data and using these values statically. Because it is not possible to exceed a certain accuracy with static values, centroid values need to be updated. Many methods have been developed for centroid updating, and this study proposes a new method. The proposed method is based on dynamically updating centroid values using the Kohonen learning rule. According to this method, centroid values are updated iteratively. In experiments conducted on the ECI CD-ROM dataset, the Kohonen learning rule was

* Sorumlu Yazar: htakci@cumhuriyet.edu.tr

observed to increase classification performance by 0.56%. This makes the Kohonen learning rule an alternative for developing centroid classifiers.

Keywords: Centroid Classifier, Model Misfit, Centroid Updating, Kohonen Learning Rule

GİRİŞ

Metinlerin içeriklerine dayalı olarak önceden belirlenmiş kategorilere atanması metin sınıflandırma olarak bilinir. Metin sınıflandırma doğal dil işleme alanının en temel problemlerinden biridir. Bugüne kadar birçok makine öğrenmesi algoritması metin sınıflandırma için kullanılmıştır (Fabrizio, 2002). Destek vektör makineleri (SVM) (Cortes ve Vapnik, 1995; Joachims, 1998), k en yakın komşu (k-NN) (Masand vd., 1992; Yang ve Liu, 1999), yapay sinir ağları (Li ve Park, 2009) ve naif bayes (Lewis ve Ringuette, 1994) bunlardan bazılarıdır. Son dönemde özellikle dönüştürücü tabanlı derin öğrenme modelleri (BERT, RoBERTa, XLNet, DeBERTa) yaygın biçimde kullanılmaktadır. Derin öğrenme modelleri bu alanda yüksek başarılar sağlamasına rağmen hesaplama maliyeti, büyük miktardaki veri ihtiyacı ve yorumlanabilirlik sorunları nedeniyle düşük maliyetli ve klasik yaklaşımlar hala önemini korumaktadır (Li vd., 2020; Wang vd, 2023; Chovanec vd., 2023).

Centroid tabanlı sınıflayıcılar bu kapsamda öne çıkmaktadır. Düşük hesaplama maliyeti ve açıklanabilirliğinin yüksek olması nedeniyle metin sınıflandırmada güçlü seçeneklerden birisidir (Han ve Karypis, 2000). Centroid sınıflayıcılar avantajlarına rağmen performans bakımından SVM sınıflayıcıların gerisindedir. Bunun nedeni; centroid tabanlı sınıflayıcıların eğitim verisine aşırı bağımlılık göstermesi (Tan, 2007) ve statik yapısıdır (Tan, 2008). Centroid değerleri güncellenmediği takdirde verideki dağılımı ve sınıflar arası ayrımları yeterli derecede sunması mümkün olamamaktadır. Bu problemin çözümü için centroid değerlerinin dinamik olarak güncellenmesine ihtiyaç vardır.

Centroid güncelleme için literatürde çeşitli çalışmalar bulunmaktadır. Bu çalışmalardan birinde düzgünleştirme ve normalizasyon teknikleri kullanılmış (Lernattee ve Theeramunkong, 2002), bir diğerinde ise sınıfları sunan centroid değerlerinin ayırt edici değerlerle ağırlıklandırılmıştır. Bu kapsamda dragpushing, hypothesis margin ve weight adjustment gibi çeşitli teknikler uygulanmıştır (Guan vd, 2009). Centroid güncelleme çalışmalarında; sınıf içi ve sınıflar arası terim frekansları kullanımının sınıflayıcı performansını artırdığı gözlenmiştir (Tan, 2008; Guan vd., 2009).

Bu çalışmada centroid güncelleme ihtiyacını karşılamak amacıyla Kohonen öğrenme kuralının kullanımı önerilmektedir. Kohonen öğrenme kuralı, benzerlik ve uzaklık ölçümüne dayalı bir öğrenme stratejisidir. Bu yönüyle centroid sınıflayıcıların uzaklık tabanlı yapısıyla yüksek düzeyde uyumluluk göstermektedir. Önerilen çalışmada Kohonen öğrenme kuralı yardımıyla klasik centroid değerleri iteratif olarak güncellenecek ve böylece güncel centroid değerleri elde edilecektir. Kohonen öğrenme kuralı centroid değerlerini veriye uydurma konusunda başarılı olabilecek bir yöntem olarak tercih edilmiştir.

Bu çalışmanın özgün katkısı, Kohonen öğrenme kuralını centroid sınıflayıcıların güncelleme mekanizması ile ilk kez bütünleştirmiş olması ve yöntemi dinamik-iteratif bir yapıya kavuşturmasıdır. Önerilen yöntem ile:

1. Centroid vektörleri klasik yöntemde olduğu gibi statik değil, iteratif biçimde güncellenmektedir,
2. Model doğru sınıflamaları ödüllendirip hatalı sınıflamaları cezalandırarak kendi kendini optimize etmektedir,
3. Yapılan çalışma ile centroid sınıflayıcıların veriye aşırı bağımlılığı (model misfit) problemi azaltılmış ve genelleme yeteneği artırılmıştır.

ECI CD-ROM veri seti üzerinde gerçekleştirilen deneylerde önerilen yöntemin klasik yaklaşıma göre doğruluğu %98,33'ten %98,89'a yükselttiği ve istatistiksel olarak anlamlı bir fark oluşturduğu belirlenmiştir. Elde edilen sonuçlar Kohonen öğrenme kuralının centroid sınıflayıcıların performansını iyileştirmede etkili bir alternatif olduğunu göstermektedir

CENTROID TABANLI SINIFLAYICI VE CENTROID GÜNCELLEME

Centroid tabanlı sınıflayıcılar sıklıkla metin sınıflandırma problemlerinde kullanılan vektör uzayı (Salton, 1989) tabanlı sınıflayıcılardır. Sınıflara ait dokümanlar ve sınıfları temsil eden değerler vektörler yardımıyla sunulmakta olup dokümanlar $\vec{d}$ ile sınıflar $\vec{c}$ ile sunulur. Dokümanları sunan doküman vektörleri ve sınıfları sunan centroid vektörleri terim sıklıklarını ya da onların tf-idf ağırlıklarını sunar. Centroid orta değer anlamına gelip kabaca bir sınıfı temsil eden eğitim dokümanlarının ortalamasından elde edilir.

Tablo 1. Çalışmada kullanılan notasyon

Sembol	Açıklama
k	Sınıf sayısı
n	Değişken sayısı
d	Doküman
$\vec{d}$	Ağırlıklandırılmış doküman vektörü
c_i	Sınıf i
$\vec{c}_i$	c_i sınıfı için centroid vektörü $\vec{c}_i = [c_{i1}, c_{i2}, \dots, c_{in}]$
t_j	terim j
D	Doküman kümesi veya derlem
D_{c_i}	c_i sınıfına ait dokümanların kümesi
C	Doküman kümesindeki sınıfların kümesi
C_{t_j}	Terim j içeren sınıfların kümesi
D_{c_i, t_j}	c_i sınıfına ait, terim j içeren dokümanların sayısı

Her sınıflandırma algoritması gibi centroid tabanlı sınıflayıcılar da eğitim ve test aşamasından meydana gelmektedir. Algoritmanın eğitim aşamasında sınıfları temsil eden centroid vektörleri elde edilir. Eğitim verisinden centroid vektörü elde etmede sıklıkla kullanılan iki yöntem vardır (Guan vd., 2009). Bunlardan birincisi aritmetik ortalama centroid (AAC) olarak bilinir. Bu yöntemde göre bir sınıf için centroid vektörü, o sınıfta yer alan doküman vektörlerinin aritmetik ortalamasından elde edilir. Bu yöntemde göre $\vec{c}_i$ vektörü aşağıdaki gibi hesaplanır:

$$\vec{c}_i = \frac{1}{|D_{c_i}|} \sum_{d \in D_{c_i}} \vec{d} \quad (1)$$

İkinci yöntem birikimli geometrik centroid (cumuli geometric centroid - CGC) olarak bilinir. Yöntem, her bir sınıf için o sınıftaki elemanların değeri toplamından bir centroid öğrenir. Bu yönetime göre $\vec{c}_i$ vektörü aşağıdaki gibi hesaplanır:

$$\vec{c}_i = \sum_{d \in D_{c_i}} \vec{d} \quad (2)$$

Literatürde AAC ve CGC yöntemlerinin değişik varyasyonları bulunmaktadır. Bunlardan birisi de sınıf özellik centroid (CFC) yöntemidir. Bu yöntem diğerlerine nispetle daha karmaşıktır. Yönteme göre her bir i sınıfındaki j terimine ait centroid değeri aşağıdaki formüle uygun olarak hesap edilir:

$$c_{ij} = b^{\frac{|D_{c_i t_j}|}{|D_{c_i}|}} \log \left(\frac{|C|}{|C_{t_j}|} \right) \quad (3)$$

Formülde yer alan b değeri bir sabit olup değeri genellikle 1'den büyüktür. Denklemin ilk parçası sınıf içi terim indeksi olarak ikinci parçası ise sınıflar arası terim indeksi olarak adlandırılır.

Centroid sınıflayıcı ile sınıflandırma veya sınıfı bilinmeyen dokümanların sınıfın tespit edilmesi benzerlik tabanlıdır. Sınıfı bilinmeyen bir doküman centroid sınıflayıcıya sunulduğunda, dokümanın sınıfını bulmak için doküman ile sınıflara ait centroid vektörleri arasındaki benzerliklere bakılır. Doküman sınıflandırmada sıklıkla kullanılan kosinüs (cosine) benzerliği centroid sınıflayıcı algoritmasında da kullanılmaktadır. Bir test dokümanı (x) ile bir c_i sınıfı arasındaki benzerlik şöyle bulunur:

$$\text{sim}(x, c_i) = \frac{\vec{x} \cdot \vec{c}_i}{\|\vec{x}\| \cdot \|\vec{c}_i\|}, \text{ for } i = 1, \dots, k \quad (4)$$

Test dokümanı ile bütün sınıflar arasındaki benzerlikler bulunduktan sonra dokümanın sınıfı için maksimum benzerliği veren sınıf aranır. Maksimum benzerliği veren centroid vektörünün sınıfı dokümanın sınıfı olarak atanır.

$$\arg \max_i \text{sim}(x, c_i)$$

KOHONEN ÖĞRENME KURALI İLE CENTROID DEĞERLERİ GÜNCELLEME

Centroid tabanlı sınıflayıcılarda centroid değeri bir kere hesaplanır ve hesaplanan değer ile sınıflandırma işlemi yapılır. Dolayısıyla hesap edilen centroid değerleri statiktir. Mevcut uygulamadaki statik yapıyı dinamik hale getirmek sınıflayıcının başarısını artırma potansiyeline sahiptir. Centroid sınıflayıcının doğru sınıflandırma yaptığı yerde ödüllendirilmesi, hatalı sınıflandırma yaptığı yerde ise cezalandırılması çözüm olarak planlanmıştır. Sınıf değerlerinin dinamik olarak yeniden hesap edilmesi ve bu işlemin yinelemeli olarak sürdürülmesi yapay

sinir ağlarındaki öğrenme kuralları ile ilişkilidir. Yapay sinir ağlarında eğer bir ağ yeteri kadar iyi sonuçlar vermiyorsa, ağın öğrenim kuralına uygun olarak ağ yeniden eğitilir ve hata oranı düşürülmeye çalışılır. YSA için birçok öğrenim kuralı bulunmasına rağmen gerek centroid sınıflayıcıya benzerlikleri gerekse kendi kendine öğrenmesi dolayısıyla kohonen öğrenme kuralı bu çalışmada tercih edilmiştir.

Kohonen öğrenme kuralının uygulandığı özellikle LVQ sınıflayıcılar centroid tabanlı sınıflayıcılar gibi benzerlik / uzaklık tabanlı sınıflayıcılardır. Bu nedenle kohonen öğrenme kuralının benzerlik / uzaklık tabanlı yöntemlere entegrasyonunun başarı şansı yüksek olacaktır.

Kohonen Öğrenme Kuralı

Kohonen öğrenme kuralı Tuevo Kohonen (1984) tarafından ortaya konmuştur. Temel felsefesi; kohonen tabakasında yer alan işlem elemanlarının birbiriyle yarışması ilkesine dayanır. Yarışma; girdi vektörü $x = (x_1, x_2, \dots, x_n)$ ile işlem elemanlarına ait referans vektörleri $w_j = (w_{1j}, w_{2j}, \dots, w_{nj})$ arasındaki öklid uzaklıklarının bulunmasına dayanır. Hangi referans vektörü girdi vektörüne daha yakınsa referans vektörü yarışmayı kazanır ve girdi vektörü yarışmayı kazanan referans vektörü ile ilgili sınıfa atanır (Öztemel, 2003).

Kohonen öğrenme kuralı algoritmik olarak ifade edilecek olursa;

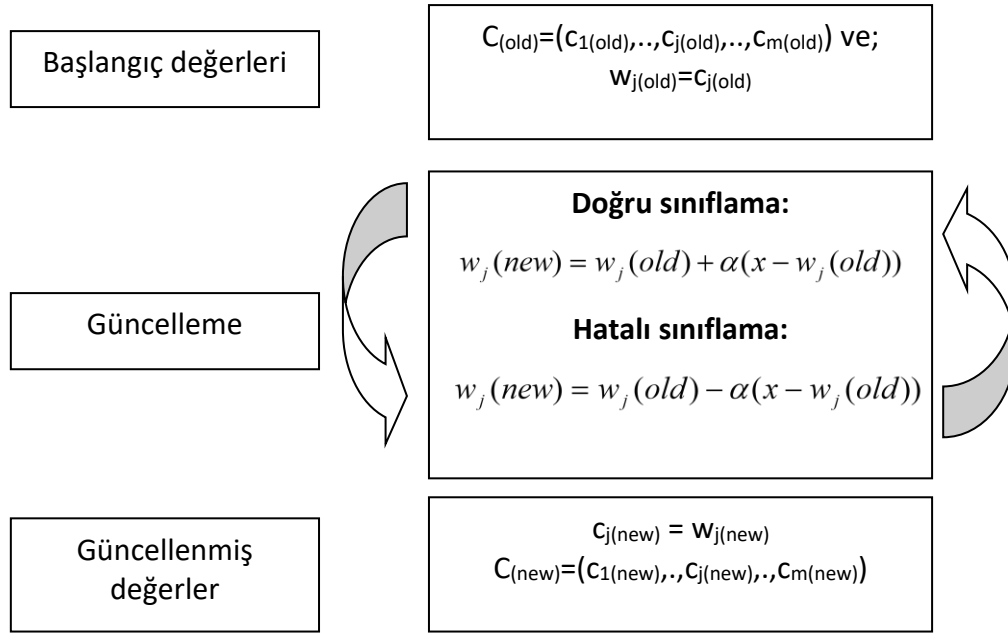
1. Girdi katmanı ile kohonen katmanı arasındaki bağlantılara ait ağırlıklara başlangıç değerleri verilir. (çalışmamızda klasik centroid yöntemi ile elde edilen değerler başlangıç ağırlıkları olarak kullanılacaktır)
2. Öğrenme kuralının durma şartı gerçekleşmediği müddetçe 3 ile 7 arasındaki işlemler yerine getirilir
3. Eğitim setinden bir örnek (x) öğrenme kuralına sunulur ve 4 ve 5. adımlar bu veri üzerinde uygulanır.
4. Öğrenme kuralına sunulan örnek ve referans vektörleri arasındaki uzaklıklar

$$Dist(j) = \sqrt{\sum_{i=1}^n (x_i - w_{ij})^2} \text{ formülü ile bulunur ve bunlardan minimum uzaklığı}$$

(Dist(j)) veren j değerini elde edilir.

5. Eğer bulunan j değeri olması gereken değer ise $w_j(new) = w_j(old) + \alpha(x - w_j(old))$ şeklinde aksi takdirde $w_j(new) = w_j(old) - \alpha(x - w_j(old))$ şeklinde ağırlık değeri güncellenir.
6. Öğrenim oranı α değeri azaltılır.
7. Durma şartına göre kontrol işlemi uygulanır ve eğer durma şartı gerçekleşir ise öğrenme işlemi son bulur. Durma şartı: hatanın belli bir değer altına düşmesi, belli sayıda yineleme veya öğrenme oranının 0 değerine ulaşması olabilir.

Kohonen öğrenme kuralının felsefesi; doğru sınıflandırmayı veren ağırlıkların artırılması, hatalı sınıflandırmayı veren ağırlıkların azaltılmasıdır. Ağırlık artırma ve azaltma için kullanılan öğrenim oranı değeri zamanla sıfır değerine kadar düşürülür. Centroid vektörlerinin güncellemesi için kohonen öğrenme kuralından faydalanırken önerilen sistemin yinelemeli yapısı aşağıda özetlenmiştir.



Şekil 1. Kohonen öğrenme kuralı yardımıyla centroid güncelleme şeması

İşlemin başlangıcında klasik centroid bulma yöntemlerinden birine göre centroid değerleri elde edilerek centroid vektörlerine sunulur. $C_{(old)} = (c_{1(old)}, \dots, c_{j(old)}, \dots, c_{m(old)})$ bu değerlere güncellenmemiş anlamında $C_{(old)}$ adı verilmiştir. Klasik centroid değerleri kohonen ağındaki giriş vektörleri ile referans vektörleri arasındaki ağırlık vektörlerine atanarak kohonen ağıının ilk değerleri $w_{j(old)} = c_{j(old)}$ şeklinde elde edilir. Bu atamaların ardından yinelemeli olarak kohonen öğrenim kuralı na göre değerler güncellenir ve güncellenen bu değerler bu sefer; $c_{j(new)} = w_{j(new)}$ ve $C_{(new)} = (c_{1(new)}, \dots, c_{j(new)}, \dots, c_{m(new)})$ şeklinde atanır. Bu atamanın ardından sınıflandırma performansı ölçülür ve en baştaki performans ile karşılaştırılır. Bu işlem durma şartına kadar devam eder.

DENEYSEL ÇALIŞMA VE ANALİZ

Önerimiz, ECI CD ROM verileri üzerinde, dile dayalı metin sınıflandırma problemi üzerinde test edilmiştir. Toplam dokuz adet dil ile yapılan çalışmada her bir dile ait ortalama 100 K eğitim ve 100 K test verisi kullanılmıştır. Deneylerde kullanılan diller şunlardır: İngilizce (eng), Fransızca (fre), Almanca (ger), Hollandaca (dut), İtalyanca (ita), Portekizce (por), Türkçe(tur), İspanyolca (spa), İsveççe (swe). Eğitim verileri; her biri 60 adet özellik değeri içeren, centroid vektörlerinin ilk değerlerinin elde edilmesi ve akabinde kohonen öğrenme kuralı ile ağırlıkların güncellenmesinde kullanılmıştır. Test verileri ise centroid değerlerinin sınıflandırma başarısını elde etmede kullanılmış ve bu veriler her biri 100 byte uzunluğunda doküman parçaları olarak işleme alınmıştır. Sınıflandırma işleminde test verilerinin ne kadarının doğru sınıfa atanıp atanmadığına göre sınıflandırma başarısı elde edilmiş olup performans parametresi olarak doğru tanıma bilgisi kullanılmıştır.

Deneysel çalışmanın başında referans amaçlı olarak önce centroid değeri ile elde edilen sınıflandırma sonucu elde edilmiş ardından kohonen öğrenme kuralı ile elde edilen centroid

değerlerinin sınıflandırma sonuçları bulunmuş ve bunlar referans sınıflandırma başarıları ile karşılaştırılmıştır.

Referans amaçlı olarak kullanılan centroid değerleri; Takcı ve Soğukpınar (2004) tarafından daha önceki bir çalışmada elde edilen yöntemle göre hesap edilmiş ve aşağıdaki sınıflandırma sonucu elde edilmiştir. Gerek ilk deneyde gerekse diğer deneylerde; her biri 100 byte uzunluğundaki dokümanlara ait 100'er adet doküman ile deneyler yapılmıştır.

Tablo 2. Güncellenmemiş centroid vektörü ile elde edilen sonuçlar

	eng	fre	ger	dut	ita	por	tur	spa	swe
eng	99	0	0	1	0	0	0	0	0
fre	0	97	0	1	0	0	0	2	0
ger	0	0	100	0	0	0	0	0	0
dut	4	0	0	96	0	0	0	0	0
ita	0	1	0	0	98	0	0	1	0
por	0	0	0	0	0	96	0	4	0
tur	0	0	0	0	0	0	100	0	0
spa	0	0	0	0	1	0	0	99	0
swe	0	0	0	0	0	0	0	0	100

Tablo 2 değerlerine bakıldığında bazı dillerin yüzde yüz oranında doğru tanıma başarısı verirken diğer bazılarının %96, %97 gibi daha düşük değerler verdiği görülür. Bizim amacımız başarılı tanımları bozmadan diğerlerinin başarısını artırmaktır. Bunun için aşamalı olarak kohonen algoritması uygulanacak ve her adımın sonunda sonuçlar karşılaştırılarak en iyi değerler bulunmaya çalışılacaktır. Kohonen öğrenme kuralı parametrelerinden biri olan öğrenim oranı 0,20 değeri ile başlatılacak ve her adımda bu değer 0,05 azaltılarak en son 0.05 değerine kadar düşürülecektir. Öğrenim oranı için 0,05 değeri son değerdir, çünkü bir sonraki aşamada öğrenim oranı 0 olacak ve güncelleme işlemi zaten son bulacaktır.

Referans centroid değerleri ile elde edilen doğru sınıflandırma oranı % 98,33 olarak elde edilmiştir. Amacımız bu doğru tanıma oranını daha yüksek bir değere taşımaktır. Kohonen öğrenme kuralı yardımıyla sınıflandırma hatasına sebep olan centroid değerleri artırılacak ve doğru sınıflandırmaya sebep olan centroid değerleri ise düşürülecektir. Öğrenim oranı 0,20 ile elde edilen sınıflandırma doğrulukları Tablo 3'de sunulduğu gibidir:

Tablo 3. Öğrenim oranı=0,20 ile güncellenmiş centroid vektörü ile elde edilen sonuçlar

	eng	fre	ger	dut	ita	por	tur	spa	swe
eng	99	0	0	1	0	0	0	0	0
fre	0	98	0	1	0	0	0	1	0
ger	0	0	100	0	0	0	0	0	0
dut	3	0	0	97	0	0	0	0	0
ita	0	1	0	0	99	0	0	0	0
por	0	0	0	1	0	98	0	1	0
tur	0	0	0	0	0	0	100	0	0
spa	0	0	0	0	4	1	0	95	0
swe	0	0	0	0	0	0	0	0	100

İlk güncelleme sonucuna göre sistemin doğruluk oranı %98,44 değerine yükselmiştir. Yukarıdaki tablo verilerine dayalı olarak ve öğrenim oranı 0,15 için sınıflandırma algoritması çalıştırılmış ve Tablo 4'deki değerler elde edilmiştir.

Tablo 4. Öğrenim oranı=0,15 ile elde edilmiş centroid vektörü ile elde edilen sonuçlar

	eng	fre	ger	dut	ita	por	tur	spa	swe
eng	99	0	0	1	0	0	0	0	0
fre	0	97	0	1	0	0	0	2	0
ger	0	0	100	0	0	0	0	0	0
dut	3	0	0	97	0	0	0	0	0
ita	0	1	0	0	97	0	0	2	0
por	0	0	0	0	0	96	0	4	0
tur	0	0	0	0	0	0	100	0	0
spa	0	0	0	0	0	0	0	100	0
swe	0	0	0	0	0	0	0	0	100

Bazı diller için sınıflandırma doğruluğu artarken diğer bazılarının sınıflandırma değeri düşmüş ve doğruluk oranı açısından bir önceki gibi %98,44 değeri elde edilmiştir. Çalışmaya devam edilmiş ve bu sefer 0,10 öğrenim oranına göre deney yapılmıştır. 0,10 öğrenim oranıyla elde edilen sınıflandırma doğrulukları Tablo 5'de görülmektedir.

Tablo 5. Öğrenim oranı=0,10 ile elde edilen centroid vektörü ile elde edilen sonuçlar

	eng	fre	ger	dut	ita	por	tur	spa	swe
eng	99	0	0	1	0	0	0	0	0
fre	0	98	0	1	0	0	0	1	0
ger	0	0	100	0	0	0	0	0	0
dut	3	0	0	97	0	0	0	0	0
ita	0	0	0	0	100	0	0	0	0
por	0	0	0	0	0	98	0	2	0
tur	0	0	0	0	0	0	100	0	0
spa	0	0	0	0	2	0	0	98	0
swe	0	0	0	0	0	0	0	0	100

Öğrenim oranı 0,10 için doğruluk oranı %98,89 olmuştur. Bu değer şu ana kadar elde edilmiş en yüksek değerdir. En son 0,05 öğrenim oranı için sonuçlar alınmış ve bu sonuçların %98,89 değerinden daha düşük olduğu görülmüştür. Kohonen algoritması ile yapılan güncelleme çalışması için en yüksek değer %98,89 olarak elde edilmiştir.

Tablo 6. Öğrenme oranı ile ilişkili performans değerleri

Öğrenme oranı	Doğruluk (%)	Kesinlik	Duyarlılık	F1
Yok	98,33	98,4	98,3	98,3
0,20	98,40	98,5	98,4	98,4
0,15	98,40	98,3	98,4	98,3
0,10	98,89	99,0	98,9	98,9

Tablo 6 değerlerine göre güncelleme yapılmayan centroid vektörleri ile elde edilen sınıflandırma doğruluğu %98,33 iken bu değer kohonen yöntemi ile güncelleme sonrası %98,89 olmuştur. Ayrıca kesinlik değeri %98,4'den %99,0'a, duyarlılık değeri %98,3'den %98,9'a ve F1 değeri %98,3'den %98,9'a çıkmıştır. Doğruluk için %0,56 birimlik artış elde edilmiştir. Diğer performans metrikleri için de artış elde edildiği görülmektedir.

Elde edilen sonuçlar istatistiksel anlamlılık açısından da analiz edilmiştir. Bunun için Paired t-testi kullanılmıştır. Paired t-testi için hipotezler şöyledir:

- H0 (Null): Tablo 2 ve Tablo 5 değerleri arasında anlamlı fark yoktur.
- H1 (Alternatif): Tablo 5 performansı anlamlı biçimde daha iyidir.

İki tablo arasındaki ortalama doğruluk farkı 0,56 ve standart sapma 0,88 ile $n=9$ iken elde edilen t değeri 1,91 ve buna karşılık gelen p değeri yaklaşık olarak 0,045 olmuştur. Elde edilen p değeri 0,05 değerinden küçük olduğu için istatistiksel olarak anlamlı (sınırdaki) bir fark olduğu tespit edilmiştir. Tablo 5 özellikle İtalyanca ve Portekizce sınıflarındaki hata oranını azaltarak toplam doğruluğu artırmıştır.

SONUÇ

Yapay sinir ağları makine öğrenmesi alanına ve sınıflandırma problemlerine oldukça önemli katkılar sunmuştur. Tek başına oldukça iyi sonuçlar veren yapay sinir ağları bu çalışmada başka bir sınıflayıcının iyileştirilmesi amacı ile kullanılmıştır. Yapay sinir ağlarında öne çıkan Kohonen öğrenme kuralı bu çalışmada centroid değerlerinin iteratif olarak güncellenmesinde kullanılmıştır. Centroid sınıflayıcılar etkili ve kolay geliştirilebilir fakat statik yapıya sahip sınıflayıcılardır. Kohonen öğrenimi sayesinde centroid sınıflayıcı zenginleştirilmiş, dinamik hale getirilmiş ve sınıflama başarısı artırılmıştır.

Bu çalışmada 0,20 öğrenim oranı ile başlayan güncelleme çalışmaları 0,05 birimlik adımlarla 0,05 öğrenim oranına kadar devam etmiş ve klasik centroid ile elde edilen %98,33 değeri %98,89 değerine çıkmıştır. Kohonen öğreniminin sağladığı kazanç %0,56 gibi önemli bir değerdir. Öğrenim oranı 0,20 yerine 0,50 gibi bir değerden başlatılıp her bir adımda 0,05 birim yerine 0,01 birim gibi bir ilerleme yapılarak daha iyi sonuçlar alınabilir.

Bu çalışmada Kohonen öğrenme kuralı kullanılarak centroid tabanlı sınıflayıcıların dinamik biçimde güncellenmesi önerilmiş ve yöntem ECI CD-ROM veri seti üzerinde başarılı sonuçlar vermiştir. Gelecekte yapılacak çalışmalarda, farklı dil veya konu alanlarına ait veri setleri üzerinde yöntemin genellenebilirliği test edilebilir. Ayrıca öğrenim oranının adaptif olarak belirlenmesi, durma koşullarının otomatik optimize edilmesi ve diğer rekabetçi öğrenme kurallarının (örneğin Oja veya Hebbian öğrenme) entegrasyonu, performansın daha da iyileştirilmesine katkı sağlayabilir. Kohonen destekli centroid güncelleme yaklaşımı, yalnızca metin sınıflandırma değil, görüntü veya ses tanıma gibi diğer örüntü tanıma alanlarına da uygulanabilir. Gelecekte bu yöntemin derin öğrenme mimarileriyle hibrit biçimde kullanılması ve GPU destekli ortamlarda hızlandırılmış sürümlerinin geliştirilmesi de önemli araştırma yönleri arasında değerlendirilebilir.

KAYNAKLAR

- Chovanec, P., et al. (2023). A survey of text classification with transformers: How wide? How large? How long? How accurate? How expensive? arXiv preprint arXiv:2311.14758.
- Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3), 273–297.
- Fabrizio, S. (2002). Machine learning in automated text categorization. *ACM Computing Surveys*, 34(1), 1–47.
- Guan, H., Zhou, J., & Guo, M. (2009). A class-feature-centroid classifier for text categorization. In *Proceedings of the WWW Conference* (pp. 1–10). Madrid, Spain.
- Han, E.-H., & Karypis, G. (2000). Centroid-based document classification: Analysis and experimental results. In *Principles of Data Mining and Knowledge Discovery* (pp. 424–431).
- Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features. In C. N. Dellec & C. Rouveirol (Eds.), *Proceedings of the 10th European Conference on Machine Learning (ECML-98)* (pp. 137–142). Springer-Verlag.
- Kohonen, T. (1984). *Self-organization and associative memory*. Springer-Verlag.
- Lertnattee, V., & Theeramunkong, T. (2002). Combining homogeneous classifiers for centroid-based text classification. In *Proceedings of ISCC* (pp. 1034–1039).
- Lewis, D., & Ringuette, M. (1994). Comparison of two learning algorithms for text categorization. In *Proceedings of the 3rd Annual Symposium on Document Analysis and Information Retrieval*.
- Li, C. H., & Park, S. C. (2009). Combination of modified BPNN algorithms and an efficient feature selection method for text categorization. *Information Processing and Management*, 45(3), 329–340.
- Li, Q., Peng, H., Li, J., Xia, C., Yang, R., Sun, L., & Yu, P. (2020). A survey on text classification: From shallow to deep learning. arXiv preprint arXiv:2008.00364.
- Masand, B., Linoff, G., & Waltz, D. (1992). Classifying news stories using memory-based reasoning. In *Proceedings of the 15th ACM International Conference on Research and Development in Information Retrieval (SIGIR-92)* (pp. 59–65). ACM.
- Öztemel, E. (2003). *Yapay sinir ağları*. Papatya Yayıncılık.
- Salton, G. (1989). *Automatic text processing: The transformation, analysis, and retrieval of information by computer*. Addison-Wesley.
- Takçı, H., & Soğukpınar, İ. (2004). Centroid-based language identification using letter feature set. In *Proceedings of CicLing* (pp. 635–645). Springer-Verlag.
- Tan, S. (2007). Large margin drag-pushing strategy for centroid text categorization. *Expert Systems with Applications*, 33(1), 215–220.
- Tan, S. (2008). An improved centroid classifier for text categorization. *Expert Systems with Applications*, 35(1–2), 279–285.
- Wang, K., Ding, Y., & Han, S. C. (2023). Graph neural networks for text classification: A survey. arXiv preprint arXiv:2304.11534.

Yang, Y., & Liu, X. (1999). A re-examination of text categorization methods. In Proceedings of the 22nd Annual International SIGIR Conference (pp. 42–49). Berkeley, CA.

Yazar Katkıları

Çalışmanın tamamı, kavramsal tasarım, veri toplama, analiz ve yazım aşaması dâhil olmak üzere yazar tarafından gerçekleştirilmiştir.

Çıkar Çatışması

Yayınla ilgili herhangi bir çıkar çatışması bulunmamaktadır.

Fonlama

Bu çalışma için herhangi bir fon desteği alınmamıştır.

Not

Makale için eklenecek not bulunmamaktadır.

Etik Bildirim

Bu çalışma için etik kurul izni gerekmemektedir.