MAKİNE ÖĞRENMESİ YARDIMIYLA OPTİK KARAKTER TANIMA SİSTEMİ
OPTICAL CHARACTER RECOGNITION SYSTEM VIA MACHINE LEARNING

Burcu BEKTAŞ
burcu.bektas@istanbul.edu.tr, İstanbul Ünv., Teknik Bilimler MYO, İstanbul
Sebahattin BABUR
sebahattin.babur@gedik.edu.tr, İstanbul Gedik Ünv., Gedik MYO, İstanbul
Uğur TURHAL
ugurturhal@balikesir.edu.tr, Balıkesir Ünv., Bilgisayar Mühendisliği, Balıkesir
Erdoğan KÖSE
erdogan.kose@istanbul.edu.tr, İstanbul Ünv., Teknik Bilimler MYO, İstanbul

ABSTRACT

Character recognition by using digital images is important for achieving the written data on paper. Machine learning system is being used for accurate and rapid estimation of characters and it is an area to be explored and developed. By using distinguishing characteristics of characters, this research aims to identify the characters of a document in an image format with a high accuracy by utilizing classification methods. Therefore, the best method has been identified by comparing the performance results and operation times of different classification methods.

Keyword: character recognition, machine learning, future extraction, classification

ÖZET

Sayısal görüntüler kullanılarak karakterlerin tanımlanması işlemi kağıt üzerindeki verilerin arşivlenmesi amacıyla önemlidir. Karakterlerin hızlı ve doğru bir şekilde tahmin edilebilmesi için makine öğrenmesi yöntemlerinden yararlanılmaktadır ve bu alan geliştirilmeye açık bir araştırma konusudur. Bu çalışmada, karakterlerin ayırtedici özelliklerinden faydalanılarak resim formatında bir belgenin, sınıflandırma yöntemleri yardımıyla karakterlerini en iyi doğruluk oranı ile tespit etmek amaçlanmaktadır. Bu nedenle farklı sınıflandırma yöntemlerinin performans sonuçları ve süreleri karşılaştırılıp en iyi yöntem belirlenmiştir.

Anahtar Kelimeler : karakter tanıma, makine öğrenmesi, öznitelik çıkarma, sınıflandırma

1. GİRİŞ

Optik Karakter Tanıma (Optical Character Recognition - OCR) sistemi, taranmış dosyalar, dijital kamerayla çekilen resimler gibi elektronik görüntüler üzerindeki karakterlerin okunarak ASCII koda dönüştürülme işlemidir. Bu yöntem, elektronik belgelerin düzenlenebilir ve aranabilir verilere dönüştürülmesine imkan sağlamaktadır. (Önal, 2013; İTÜ Bilgi İşlem, 2013). OCR sistemi, görüntüdeki harfleri seçip ayırarak, harflerden kelimeleri, kelimelerden de cümleleri oluşturmaya ve bu sayede metnin ayıklanmasına olanak tanır.

Karakter tanıma teknolojilerinin sağladığı kolaylıklar, birçok iş alanında bu teknolojilerin hızla uygulanmaya başlanmasına yardımcı olmuştur (Şekerci, 2007). Örneğin; mektupların, üzerinde bulunan posta koduna göre ayrıştırılması, bankalara yollanan çeklerin otomatik olarak tanınıp gerekli hesap işlemlerinin elektronik ortamlarda gerçekleştirilmesi, kütüphanelerdeki kitapların bilgisayar ortamına aktarılması, otoparklarda,

geçiş kontrolünün olduğu alanlarda ve plaka tanıma sistemlerinde, reklam, afiş, market panolarının okunması gibi projelerde karakter tanıma teknolojileri yaygın olarak kullanılmaktadır (Bektaş, 2014).

Bu konuda farklı yöntemlerle gerçekleştirilen birçok çalışma ele alınmıştır. Berrin Yanıkoğlu'nun yaptığı çalışmada harflerin bölümlenmesi üzerinde durulmuştur ve %97 doğruluk oranı elde edilmiştir (Yanıkoğlu & Sandon, 1998).

Jianchang Mao ve arkadaşlarının yaptığı çalışmada el yazısıyla yazılmış adres tanıma sistemi geliştirilip %83.5 doğruluk elde edilmiştir (Mao, Sinha, & Mohiuddin, 1998). Ni ise Amerikan posta servislerindeki adresleri sıralamak için posta kodlarını okuyan çok katmanlı geri-beslemeli yapay sinir ağı kullanan bir OCR sistemi geliştirmiştir (Ni, 2007).

C. Ng ve arkadaşının yaptığı çalışmada Braille (görme engelli insanların okuyup yazması için kullanılan bir alfabe yöntemi) sayfalarını tanımak amaçlanmıştır ve %97 doğruluk oranı elde edilmiştir (Ng & Vincent Ng, 1999).

Campos ve arkadaşlarının yaptığı çalışmada bazı sokak resimlerinden elde edilen görüntülerdeki karakterleri tanımak amaçlanmıştır. (Campos, Babu, & Varma, 2009).

Halit Çetiner ve arkadaşlarının çalışmasında, Türkiye Cumhuriyeti (TC) kimlik numaralarının kamerayla çok kısa zamanda tespiti ve veri tabanından kişi bilgilerinin çağrılması işleminin gerçek zamanlı olarak yapılması amaçlanmıştır (Çetiner, Cetişli, & Çetiner, 2012).

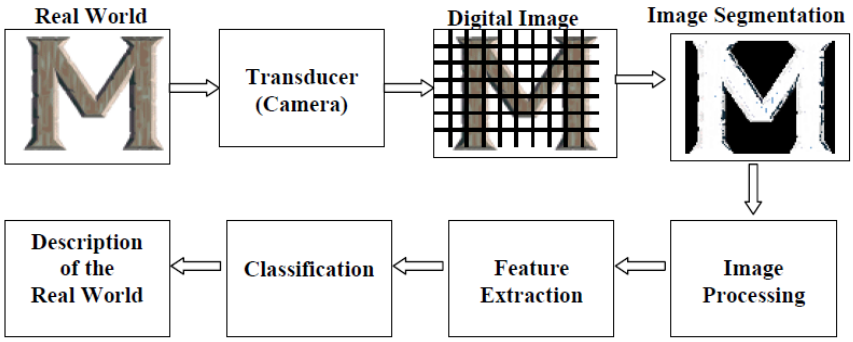
Gorski ve arkadaşları, Fransa, İngiltere veya Amerika'da düzenlenen el yazısı veya basılı çekleri işlemek için bir çek tanıma sistemi gerçekleştirmiştir (N.Gorski, ve diğerleri, 1999).

Leung ve daha sonra Zhang ve arkadaşları 2002 yılında, el yazısı Çin karakterlerini tanıyan çalışmalar geliştirmiştir (Sze, 1997; Zhang, Zhao, Yang, & Wang, 2002).

Bu çalışmada The Chars74K veri setinde bulunan 36 farklı karakter örüntüsünden (0-9,A-Z) öznitelikler çıkarılmış ve bu öznitelikler çeşitli algoritmalar kullanılarak sınıflandırılmıştır (The Chars74k). Elde edilen sınıflandırma sonuçları Doğruluk, Kappa, Hesaplama Süreleri ve ROC (Reciever Operating Characteristics – Alıcı İşlem Karakteristiği) değerleri ile karşılaştırılarak performans analizi gerçekleştirilmiştir. Performans açısından en iyi yöntemin belirlenmesinde bu 4 parametre birlikte değerlendirilmiştir.

2. METARYAL VE YÖNTEM

Karakter Tanıma Sistemleri temel olarak alfanümerik veriler üzerinde işlem yapan bir örüntü tanıma sistemidir. Tanıma işlemi birkaç ara aşamadan meydana gelmektedir. Her kademenin kendine ait sorun çözme yöntemi vardır ve bunlardan herhangi birinin yaklaşımın hatalı olması karakterlerin tanınma başarısını azaltmaktadır. Örnek bir karakter tanıma sistemi Şekil-1' de gösterilmektedir. (Ovatman, 2005).



Şekil 1: Örüntü tanıma alt aşamaları

Transducer; genellikle bir kamera, tarayıcı şeklindedir ve görüntünün sayısallaştırılarak bilgisayar ortamında işlenebilir hale getirilmesine yönelik bir aşamadır. Prosessing (ön işleme) kademesi; görüntünün

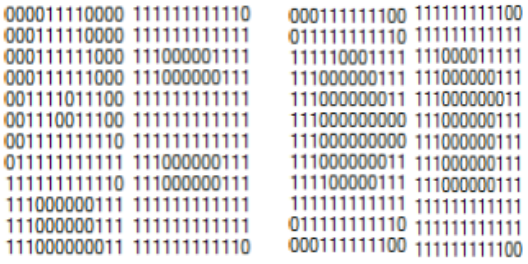
normalleştirilmesi amacıyla yapılan yumuşatma ve filtreleme gibi ön işlemlerden oluşmaktadır. Feature Extraction (öznitelik çıkarma); işlenecek olan verinin sahip olduğu sınıfa ait ortak özellikleri çıkarma işlemidir. Classification (sınıflandırma); söz konusu bir desenin sınıf üyeliğine ilişkin karar verme sürecidir. Bir karakter tanıma sisteminin temel amacı; tüm bu süreçler için farklı yöntemler kullanılarak en iyi doğruluk oranının elde edilmesidir.

Bu çalışmada aşağıda örnek gösterimleri bulunan 36 adet karakterin her birinden alınan 100 örnek üzerinde öznitelik çıkarma ve sınıflandırma aşamaları üzerinde durulmuştur.

Şekil 2 : Kullanılan veri setine ait örnek karakter kümesi

Öznitelik Çıkarma

Şekil 1 ve Şekil 2'de verilen optik karakterlerin ve bu karakterleri tanımlayacak sistemin oluşturulması için öncelikle giriş örüntülerinden özniteliklerin çıkarılması gerekmektedir. Bu işlem için kullanılan birçok yöntem mevcuttur. Bu çalışmada; giriş örüntülerine otomatik eşikleme yapılarak ikili resimler elde edilmiştir. Otomatik eşikleme yönteminde 0-255 arasında ifade edilen siyah-beyaz giriş örüntüsünde piksel değeri 127' den düşük olanlar 1, eşit ve yüksek olanlar ise 0 olarak etiketlenmiştir. Böylece sayısallaştırılan karakterlerin gözle görülür seviyede ayrıştırılması sağlanmıştır. Şekil 3'te; bu yöntemin uygulandığı örnek A, B, C ve D karakterlerinin ikili matris formatında gösterimi bulunmaktadır.



Şekil 3: Karakterlerin matris formatında gösterimi

Otomatik eşikleme algoritmasından geçirilen giriş örüntülerinde; örüntünün orijini referans alınarak, saat yönünde 45'er derecelik açılarla -dıştan içe- 1 karakteri aranmış, elde edilen koordinat değeri Mesafe Hesaplama Algoritması (MHA) yardımıyla uzaklığa dönüştürülmüştür. MHA Eşitlik-1 ile ifade edilmiştir;

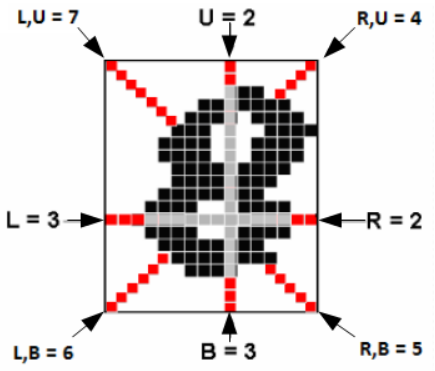
$$MHA = \sqrt{{i_x}^2 + {j_y}^2} \tag{1}$$

Eşitlikte;

i_x: Tespit edilen karakter sınırının X eksenindeki değerini,

j_y: Tespit edilen karakter sınırının Y eksenindeki değerini temsil etmektedir.

Bu algoritma sonucunda tüm karakterler için 1x8' lik bir vektör oluşmaktadır. Öznitelik çıkarma işleminde kullanılan MHA algoritmasının çalışma yöntemi Şekil 4'te gösterilmiştir.



Şekil 4: MHA algoritması ile özniteliklerin elde edilmesi

Sınıflandırma

Bir önceki aşamada öznitelikleri oluşturulan giriş örüntüleri için sınıflandırma işlemleri aşağıda belirtilen algoritmalar yardımıyla gerçekleştirilmiştir:

- Naïve Bayes
- k-NN (Öklid Mesafe Fonksiyonu)
- LibSVM (Lineer Çekirdek Fonksiyonu)

Naive Bayes algoritmasında sınıflandırma için normalize edilmiş öznitelik vektörleri, sınıflandırıcı uzayında kullanılır. Bu algoritma; verilen örneklerin en büyük olasılıklarını hesaplama problemi olarak görülebilir. Sınıflandırma modeline ihtiyaç duymayan ve olasılık temelli çalışan algoritmada, farklı problem tiplerine göre düşük hesaplama sürelerinde oldukça verimli sonuçlar elde edilmektedir.

k-En Yakın Komşuluk (k-NN) algoritması, kümeleme problemini çözen en temel gözetimsiz öğrenme yöntemleri arasında yer alır. Algoritmanın genel mantığı; n adet veriden oluşan bir veri kümesini, giriş parametresi olarak verilen k ($k < n$) adet kümeye bölümlemektir. Amaç, gerçekleştirilen bölümleme işlemi sonunda elde edilen kümelerin, küme içi benzerliklerinin maksimum ve kümeler arası benzerliklerinin minimum olmasını sağlamaktır. Yöntemin performansını k küme sayısı, başlangıç olarak seçilen küme merkezlerinin değerleri ve benzerlik ölçümü kriterleri etkilemektedir (Zouhal & Denceux, 1998). Bu çalışmada k değeri 1 olarak belirlenmiştir.

LibSVM, Destek Vektör Makinesi (SVM) için geliştirilmiş ve en yaygın kullanılan SVM kütüphanelerinden biridir. İki sınıflı verilerin tahmininde güçlü bir makine öğrenme tekniği olan geleneksel SVM'nin aksine çok sınıflı verilerin kullanılmasına olanak sağlamaktadır. LibSVM çalışması iki aşamadan oluşmaktadır: İlk aşamada model oluşturmak için veri seti eğitilir, ikinci aşamada oluşturulan model test veri setine ait bilgilerin tahmini için kullanılır (Chang & Lin, 2011; Bektaş & Babur, 2016). LibSVM; sınıflandırma, regresyon ve dağılım tahmini için SVM fonksiyonlarını kullanır. Bu çalışmada LibSVM fonksiyonu, SVM sınıflandırıcısının Lineer çekirdeği kullanılarak oluşturulmuştur.

N Kümeli Çapraz Doğrulama metodu; veri madenciliği çalışmalarında, elde edilen sonuçların geçerliliğini arttırmak amacıyla kullanılmaktadır. Veri kümesi eğitim ve test olarak ayrıştırılırken; her birinde eşit sayıda farklı veri türü içeren N adet küme oluşturur ve bu kümelerden N-1 adedinin eğitim, kalanının da test kümesi olarak kullanılmasını sağlar.

Bu çalışmada kullanılan 3600 örnekli veri seti için N değeri 10 olarak seçilmiştir. Bu durumda veri kümesi, her biri 360 örnek içeren 10 eşit kümeye bölünmüş, bu bölümleme sonucunda 3240 örnek eğitim, 360 örnek ise test amaçlı kullanılmıştır. Bu işlem, her kümenin test olarak kullanılması amacıyla 10 kez tekrarlanmış ve her sınıflandırma sonucunda elde edilen doğruluk, kappa, ROC ve hesaplama sürelerinin ortalamaları alınarak algoritmanın performansı belirlenmiştir.

3.BULGULAR

Çalışmada kullanılan ve Şekil 2'de sunulan veri seti 0-9 arası sayısal karakterlerden ve ingiliz alfabesinde bulunan harflerin büyük hallerinden oluşmaktadır. 36 karakter ve her karaktere ait 100 örnekten Mesafe Hesaplama Algoritması kullanılarak 8'er adet öznitelik çıkartılmıştır. Elde edilen öznitelikler; Naive Bayes, Lineer çekirdek fonksiyonlu LibSVM ve Öklid bağıntılı k-NN algoritmaları yardımıyla, 10-Kat Çapraz Doğrulama (Ten Fold Cross Validation) – Birini Dışarıda Bırak (Leave One Out, LOO) yöntemi kullanılarak sınıflandırılmış ve elde edilen sonuçlar Tablo 1'de gösterilmiştir.

Tablo 1: Sınıflandırma Performans Sonuçları

Sınıflandırıcı	Performans Değerleri			
	Doğrıluk(%)	Kappa	ROC	Hesaplama Süresi
Naive Bayes	65.0278	0.640	0,975	1,33sn
LibSVM (Linear Kernel)	81.5278	0.810	0,905	40,02sn
k-NN (k=1) (Euclidean)	89.2778	0.889	0,953	1,65sn

Tablo 1'de verilen doğruluk değeri, doğru sınıflandırılmış örnek sayısının toplam örnek sayısına oranını, ROC (Reciever Operating Characteristics – Alıcı İşlem Karakteristiği) değeri sınıflandırma sonucunun şans ile olan ilişkisini, Kappa katsayısı ise elde edilen sonuçlar arasındaki uyuşmayı belirtir. Performans analizi yapılırken bu üç kriterin yanında hesaplama süresi de göz önüne alınmıştır. Doğruluk değerinin 100'e, ROC ve Kappa değerlerinin ise 1'e yakınlığı, sınıflandırma işleminin kalitesini ve geçerliliğini göstermektedir. Verilen hesaplama süreleri, sınıflandırıcıların; model oluşturma ve 10 kat çapraz doğrulama işlemlerinin tümünü tamamlayıncaya kadar geçen zamandır (Turhal & Akbaş, 2016).

4.TARTIŞMA, SONUÇ VE ÖNERİLER

Çok sınıflı verilerin sınıflandırılması algoritmalar için oldukça zorlu bir problemdir. Özellikle bu tip karakter tanıma çalışmalarında her karakter grubu için bir sınıf değerinin üretilmesi ve karakterlerin birbirlerine olan benzerliği, sınıflandırma sonuçlarını olumsuz yönde etkilemektedir.

Tablo 1'de elde edilen sonuçlar incelendiğinde; en düşük hesaplama süresine sahip Naïve Bayes algoritmasının aynı zamanda en düşük doğruluk değerleri ürettiği görülmektedir.

Güçlü bir ikili sınıflandırma algoritması olan Destek Vektör Makineleri (SVM)' nin güncellenmiş versiyonlu LibSVM algoritması, Naïve Bayes algoritmasına oranla daha yüksek doğruluk değerleri üretmiş olmasına rağmen, uzun hesaplama süresi bu algoritmanın kullanışlılığını azaltmaktadır.

Naïve Bayes algoritmasına çok yakın hesaplama sürelerine sahip k-NN algoritmasının herhangi bir sınıflandırma modeline ihtiyaç duymadan en yüksek doğruluk ve Kappa değerlerini üretmesi, problemin çözümünde bu algoritmanın oldukça etkili olduğunu ortaya koymaktadır. Basit matematiksel ifadeler ile hesaplanması ve yüksek performans değerleri üretmesi, bu algoritmanın düşük maliyetli ve taşınabilir bir sistemde kullanılabilirliğini arttırmaktadır.

Bundan sonraki çalışmalarda farklı öznitelik çıkarma algoritmalarının birlikte kullanılmasıyla oluşturulacak hibrit bir model, sınıflandırma sonuçlarında kalitenin arttırılmasını sağlayabileceği gibi oluşturulan hibrit modele uygulanacak öznitelik indirgeme algoritmaları, veri setini temsil edecek minimum sayıdaki özniteliklerin belirlenmesinde ve uzun hesaplama sürelerinin azaltılmasında etkili olabilir.

5.KAYNAKÇA

Bektaş, B. (2014). RFID ve XBEE Tabanlı Depo Yönetim Sistemi Tasarımı ve Gerçekleştirilmesi. İstanbul: Marmara Üniversitesi, Fen Bilimleri Enstitüsü, Yüksek Lisans Tezi.

Bektaş, B., & Babur, S. (2016). Machine Learning Based Performance Development for Diagnosis of Breast Cancer. *Medical Technologies Conference (TıpTekno)*. Antalya.

Campos, T. E., Babu, B. R., & Varma, M. (2009). Character Recognation in Natural Images. *Character Recognition in Natural Images*, (s. 273-280).

Chang, C.-C., & Lin, C.-J. (2011). LIBSVM: A Library for Support Vector Machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*.

Çetiner, H., Cetişli, B., & Çetiner, İ. (2012). Gerçek Zamanlı T.C. Kimlik Numarası Tanıma. *Sakarya Üniversitesi Fen Bilimleri Dergisi*, 123-129.

Feng, P.-M., Ding, H., Chen, W., & Lin, H. (2013). Naive Bayes Classifier with Feature Selection to Identify Phage Virion Proteins. *Computational and Mathematical Methods in Medicine*.

İTÜ Bilgi İşlem, D. B. (2013, 9 8). http://bidb.itu.edu.tr/seyirdefteri/blog/2013/09/08/ocr-(optical-character-recognition---optik-karakter-tan%C4%B1ma) adresinden alındı

Mao, J., Sinha, P., & Mohiuddin, K. (1998). A System for Cursive Handwritten Address Recognition. *Pattern Recognition, Proceedings. Fourteenth International Conference on.* (s. 1285-1287). IEEE.

N.Gorski, V.Anisimov, E.Augustin, O.Baret, D.Price, & J.-C.Simon. (1999). A2iA Check Reader: A Family of Bank Check Recognition Systems. *Document Analysis and Recognition, ICDAR'99. Proceedings of the Fifth International Conference on* (s. 523-526). IEEE.

Ng, C. M., & Vincent Ng, Y. L. (1999). Regular Feature Extraction for Recognition of Braille. *Third International Conference on Computational Intelligence and Multimedia Applications*, (s. 302-306).

Ni, D. X. (2007). Application of Neural Networks to Character Recognition . *Proceedings of students/faculty research day, CSIS, Pace University, May 4th.*

Ovatman, T. (2005). *A Real-Time Optical Character Recognition System.* İstanbul: İstanbul Technical University, Institute of Science and Technology Master Thesis.

Önal, M. (2013). *Gömülü Sistemler ile RFID Mimarisi ve Programlama.* İstanbul: KODLAB Yayın Dağıtım Yazılım ve Eğitim Hizmetleri San. ve Tic. Ltd. Şti.

Sze, C. H. (1997). Feature Selection in the Recognition of Handwritten Chinese Characters. *Engineering Applications of Artificial Intelligence*, 495-502.

Şekerci, M. (2007). Birleşik ve eğik Türkçe el yazısı tanıma sistemi. *Trakya Üniversitesi, Fen Bilimleri Enstitüsü, Bilgisayar Mühendisliği Bölümü*. Yüksek Lisans Tezi.

The Chars74k, d. (tarih yok). *The Chars74K dataset.* http://www.ee.surrey.ac.uk/CVSSP/demos/chars74k/ adresinden alındı

Turhal, U., & Akbaş, A. (2016). Yüz Örüntülerinden Cinsiyet Tespitinde Hibrit Özniteliklerle Performans Geliştirme. *Elektrik-Elektronik ve Bilgisayar Sempozyumu.* Tokat.

Yanıkoğlu, B., & Sandon, P. A. (1998). Segmentation og OFF-LINE Cursive Handwriting Using Linear Programming. *Pattern Recognition*, 1825-1833.

Zhang, L.-X., Zhao, Y.-N., Yang, Z.-H., & Wang, J.-X. (2002). Feature selection in recognition of handwritten Chinese characters. *Machine Learning and Cybernetics, Proceedings. 2002 International Conference on* (s. 1158-1162). IEEE.

Zouhal, L. M., & Denceux, T. (1998). *An evidence-theoretic k-NN rule with parameter optimization.* IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews).