

Parkinson Hastalığının Rotasyon Ormanı Yöntemi ile Teşhisi

*¹Ayşegül Sağlam, ²Ayşe Betül Gezeroğlu, ³Elif Şeflek ve ⁴Murat Gök
^{1,2,3,4}Bilgisayar Mühendisliği Bölümü, Yalova Üniversitesi, Yalova, Türkiye

Özet

Nörolojik bozukluklardan kaynaklanan Parkinson hastalığı, ruhsal ve zihinsel bozukluklar oluşturur. Hastalığın tanısı klinik muayene ile konulabilir. Bunun için hastanın muayene bulguları, tıbbi geçmişleri, özellikle belirli hareketleri yapış tarzları dikkate alınır. Parkinson hastalığının uzaktan gözetim ile teşhisinde konuşma şeklinin analizi uygulamaları giderek artmaktadır. Bu nedenle biz, Parkinson hastalığı olan kişilerin kelime ve cümleleriyle derlenmiş konuşma egzersizlerinin de bulunduğu bir veri seti kullandık. Amacımız insanların Parkinson hastası olup olmadıklarını, bilinen makine öğrenmesi tekniklerini kullanarak belirlemektir.

Anahtar sözcükler: Parkinson hastalığı, Sınıflandırıcı algoritmalar, Performans metrikleri.

Diagnosis of Parkinson's Disease Using Rotation Forest

Abstract

Parkinson's disease which is caused by neurological disorders, creates emotional and mental disorders. The diagnosis can be confirmed by clinical examination so attention will be paid to patient's examination, medical history, especially certain movements stick style .Therefore we used a dataset which has Parkinson disease's words, phrases and speaking exercises. In this study, we aimed to determine using conventional machine learning techniques that people have or do not have Parkinson's disease.

Key words: Parkinson disease, Classification algorithms, Performance metrics

1. Giriş

Parkinson hastalığı, hareket ve koordinasyonumuzu sağlayan beyin hücrelerinin bir kısmının çeşitli sebeplere bağlı zedelenmesi veya ölmesi sonucu ortaya çıkan bir hastalıktır. Parkinson hastalığı, Alzheimer hastalığından sonra en sık görülen nörodejeneratif hastalıktır. Son yıllarda genetik formlarının ve patojenik mutasyonların tanımlanması ile Parkinson hastalığının nedenlerine ilişkin bilgiler artmıştır [1]. Parkinson hastalığı toplumda kadınlara oranla erkeklerde daha fazla görülmektedir. Hastalığın ilk belirtisi titremedir. Titreme dinlenirken olup, uyurken kaybolur; sinirlilik ve yorgunluk titremeyi artırır. Yazı yazma da koordinasyon sağlanamaz örneğin yazıya büyük başlar, harfler gittikçe küçülür ve yazının okunması güçleşir. Mimik ve jestler silik, donuk ve anlamsızdır. Hareketlere başlamakta güçlük çeker ve hastaların %40'ında zamanla bunama görülür [2, 3].

Bugüne kadar yapılan çalışmalarda, Parkinson hastalığının teşhisinde pek çok yöntem önerilmiştir. Mandal ve Sairam yaptıkları çalışmada destek vektör makinaları (DVM) algoritmasını kullanmıştır [4]. Shahbakhhi ve arkadaşlarının yaptığı çalışmada, ses analizine göre hastalığın teşhisini geliştirilmişler ve DVM kullanmışlardır [5]. Daş, Parkinson hastalarını teşhis etmek için Yapay Sinir Ağı (YSA) temelli sınıflandırıcı kullanmıştır [6]. Little ve arkadaşları, DVM sınıflandırıcı ve Süzgeç Temelli İlişkisiz Ölçme (STİÖ) ile öznelikleri seçerek Parkinson hastalığı veri setindeki hasta ve normal kişileri sınıflandırmıştır [7]. Sakar ve Kursun sınıflandırma algoritmalarından lineer DVM kullanmışlardır [8]. Gök ise yaptığı çalışma da k-En Yakın Komşuluk (k-EYK) algoritmasını kullanmıştır [9].

Bu çalışmanın amacı, Parkinson hastalığının, biyomedikal ses parametrelerine dayanan bir veri seti üzerinde makine öğrenmesi yöntemleriyle teşhisinin araştırılmasıdır.

2. Materyal ve Yöntemler

2.1 Veri Seti

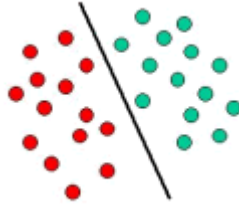
Bu çalışmada kullanılan Parkinson hastalığı veri seti, İstanbul Üniversitesi Cerrahpaşa Tıp fakültesi Nöroloji Anabilim Dalında oluşturulmuştur. Eğitim verileri 20 PWP (6 Kadın, 14 Erkek) ve 20 sağlıklı birey (10 Kadın, 10 Erkek) den oluşur. Tüm deneklerden ses kayıtlarının birden fazla türleri (26 örneğe sayılar, sözcükler ve kısa cümlelerde dahil olmak üzere) alınmıştır. Veri seti 16 öznelik ve 850 örnekten oluşmaktadır [8].

2.2 Sınıflandırma Algoritmaları

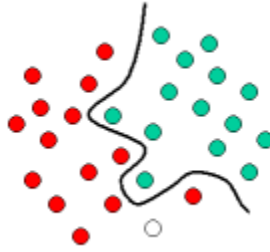
Bu çalışmada DVM, k-en yakın komşuluğu (k-EYK), Rotasyon Ormanı (RO), Naive Bayes (NB) ve C4.5 sınıflandırma algoritmaları kullanılmıştır.

DVM istatistiğe bağlı öğrenme teorisiyle oluşan kontrollü sınıflandırma algoritmasıdır. Şekil 1'de DVM'nin iki sınıflı doğrusal verilerin sınıflandırılması gösterilmiştir. Şekil 2'de çok sınıflı, doğrusal olmayan verilerin sınıflandırılması genelleştirilmiştir [10]. DVM, Cortes ve Vapnik tarafından ortaya atılmıştır ve yapısal risk minimizasyonunu kullanmaktadır. Bu yöntemde, iki sınıf arasındaki birbirine en yakın örneklerin uzaklıklarının maksimum hale getirildiği yüksek bir

düzlem bulunması amaçlanır. Doğrusal olarak ayırlamayan veriler için, DVM yardımıyla giriş vektörü yüksek boyutlu bir uzaya doğrusal olmayan bir fonksiyon yardımıyla eşleştirilir [11].



Şekil 1: Doğrusal DVM



Şekil 2: Doğrusal olmayan DVM

NB makine öğrenmesi ve veri madenciliği için en verimli ve etkili öğrenme algoritmalarından biridir. NB sınıflandırma yönteminde, öznelik vektöründeki özneliklerin tamamının bağımsız olduğu kabul edilir [12]. NB sınıflandırıcı belirli bir sınıfa ait her bir örneğin olasılığını bulmak için Bayes istatistik ve Bayes teoremi kullanır. NB sınıflandırıcı belirli bir sınıfa ait her örneğin o sınıfa ait olasılığını bulur [13]. NB sınıflandırıcı da iki önemli kabul yapılmaktadır.

- Niteliklerin hepsi aynı seviyede önemlidir,
- Niteliklerin birbirinden bağımsız oldukları kabul edilmiştir [14].

NB, Bayes teoreminin bağımsızlık önermesiyle sadeleştirilmiş halidir. Bayes teoreminin denklemi aşağıdaki gibidir;

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \quad (1)$$

Burada;

$P(A|B)$: B olayı gerçekleştiği durumda A olayının gerçekleşme olasılığıdır.

$P(B|A)$: A olayı gerçekleştiği durumda B olayının gerçekleşme olasılığıdır.

$P(A)$ ve $P(B)$: A ve B olaylarının önsel olasılıklarıdır.

$x \in \mathbb{R}^L$ ve S ayrıştırılacak sınıflar kümesiye, Bayes teoremine göre aşağıdaki ifade yazılır;

$$P(S_i \setminus x) \times P(x) = P(x \setminus S_i) \times P(S_i) \quad (2)$$

$$p(x) = \sum_{i=1}^L p(x \setminus S_i) \times P(S_i) \quad (3)$$

Burada;

$P(S_i)$: S_i 'nin öncel olasılığı $i = 1, 2, \dots, L$, $P(S_i|x)$: S_i 'nin ardıl olasılığı

$p(x)$: x in Olasılık yoğunluk fonksiyonu

$p(x|S_i)$: $i = 1 = 2, \dots, L$, x 'in koşullu Olasılık yoğunluk fonksiyonu.

Sınıfı belli olmayan bir veri olduğu durumlarda, eğer;

$P(S_i \setminus x) > P(S_j \setminus x)$, $\forall j \neq i$ ise, bayes karar teorisine göre x , sınıf S_i 'ye aittir.

Yukarıda formülle elde edilen Bayes karar teoreminde istatistik olarak bağımsızlık önermesinden faydalanılırsa bu tip sınıflandırmaya NB sınıflandırılması denir. Yani ;

$$P(S_i) \prod_{k=1}^L P(x_k \setminus S_i) > P(S_j) \prod_{k=1}^L P(x_k \setminus S_j) \quad (4)$$

RO, Sınıflandırılması veri kümesinden X tane alt küme oluşturulur. Bu X tane alt kümeye Temel Bileşen Analizi (TBA) uygulanarak boyut daraltma işlemi gerçekleştirilir. Böylece ayırt ediciliği yüksek olan özellikler seçilmiş olur. Rotasyon ormanlarında, Temel olarak bootstrap algoritması kullanılır [16,17]. RO algoritması, her ağaç için farklı öznitelikler kullanarak, bağımsız R karar ağaçları yetiştiren bir topluluk yöntemidir. Her ağaç, önce farklı bileşen analizi (Principal Component Analysis, PCA) kullanılarak eğitilir. Bunun için veri kümesinin bir alt kümesi rastgele seçilerek kullanılır.

K -en yakın komşuluk algoritmasının amacı, eğitim seti içerisindeki k en yakın komşuları arasında görünmeyen bir noktaya baskın sınıf ataması yapmaktır [18]. Yöntemin adımları şu şekilde gerçekleştirilir; İlk önce test kümesi k tane alt kümeye bölünür ve rastgele merkezler atanır. Daha sonra her noktanın merkezine uzaklıkları hesaplanır ve her veri kendisine daha yakın olan merkezin kümesine atanır. Merkez noktalar değişmeye kadar aynı işlem tekrarlanır.

C4.5 karar ağacı oluşturma algoritmasıdır. ID3 algoritmasının eksik yönlerini gidermek için Quinlan tarafından geliştirilmiştir. Sayısal değerlere sahip niteliklerin de karar ağaçlarını oluşturma olanağını sağlamıştır. Ağaçtaki her bir düğüm, özelliği, her bir dal, özellik değer bilgisini ve her bir yaprak düğüm ise sınıf etiketini belirtmektedir. Ağaç yapıları kurallarla ve sade olarak tanımlanabildiği için karar ağaçları yaygın olarak kullanılır. Bu şekilde belirlenen kurallar kolay bir şekilde iletilmiş olur [16].

3. Deneysel Sonuçlar

Parkinson hastalığının teşhisi için yapılan deneysel çalışma da sınıflandırıcı algoritmalar 10-kat çapraz doğrulama (ÇD) test tekniğini ile uygulanmıştır. 10-kat ÇD'de veriler 10 ayrı kümeye ayrılır ve 1'i test 9'u eğitim için kullanılır. Testler sonucunda sınıf doğruluğu, MCC, F-Score ve Eğri Altında Kalan Alan (EAKA) performans metrikleri elde edilmiştir. Performans metrikleri Tablo 1'de görülen karmaşıklık matrisinden elde edilir.

Tablo 1: Karmaşıklık Matrisi

		Gerçek	
		Pozitif	Negatif
Tahmin	Pozitif	Doğru Pozitif (DP)	Yanlış Pozitif (YP)
	Negatif	Yanlış Negatif (YN)	Doğru Negatif (DN)

Performans metrikleri;

$$\text{Sınıf doğruluğu: } \frac{DP+DN}{DP+DN+YP+YN} \quad (5)$$

$$\text{Duyarlılık} : \frac{DP}{DP+YN} \quad (6)$$

$$\text{F-Skor} : \frac{2DP}{2DP+YP+YN} \quad (7)$$

$$\text{ROC Alanı} : 1-\text{özgünlük}\left(\frac{DN}{YP+DN}\right) \quad (8)$$

Tablo2: Deneysel kullanılan sınıflandırma algoritmaları ve performans metrikleri

	Sınıf Doğruluğu (%)	Duyarlılık	F-skor	ROC
NB	83.1126	0.831	0.83	0.901
k-EYK	93.7086	0.937	0.937	0.938
RO	99.0894	0.991	0.991	0.991
L-DVM	96.4404	0.964	0.964	0.966
RBF-DVM	56.9536	0.57	0.413	0.5
C4.5	98.5927	0.986	0.986	0.986

Deneysel çalışmalarda RO algoritması % 99,5 sınıf doğruluğu ve % 99,10 duyarlık değerleri ile en yüksek performansı sergilemiştir. Buna karşın RBF-DVM Algoritması %56,95 değeri ile en düşük sınıf doğruluğu değerini vermiştir.

Sakar ve arkadaşları bu çalışma da kullanılan Parkinson hastalığı veri seti üzerinde % 85 sınıf doğruluğu ve % 80 duyarlık değerleri elde etmişlerdir [8].

Sonuçlar

Biz bu çalışma da Parkinson hastalığının teşhisini in silico ortamda makine öğrenmesi yöntemleri ile gerçekleştirdik. Elde edilen sonuçlar klinik ortamlarda uygulanması açısından umut vericidir. Gelecekte yapılacak çalışmalarda farklı veri setleri üzerinde sınıflandırıcı algoritmaların birleştirilmesi uygulamaları yapılması planlanmaktadır.

Teşekkür

Bu çalışma, Yalova Üniversitesi, Yüksek Lisans Projesi (2014/YL/37) tarafından desteklenmiştir.

Kaynaklar

- [1] Aslan H. Parkinson hastalığı tanısı alan hastalarda Parkinson hastalığından sorumlu olduğu bildirilen SNCA ve LRRK2 genlerindeki A30P ve G2019S mutasyonlarının araştırılması. Eskişehir Osmangazi Üniversitesi / Tıp Fakültesi / Tıbbi Genetik Anabilim Dalı, 2010.
- [2] Aarsland D. et al. Mental symptoms in Parkinson's disease are important contributors to caregiver distress. International journal of geriatric psychiatry, 14.10 1999, p. 866-874.
- [3] Edwards L. et al. Gastrointestinal symptoms in parkinson disease: 18-month follow-up study. Movement disorders, 8.1 1993 p. 83-86.
- [4] Mandal I, Sairam N. Accurate telemonitoring of parkinson's disease diagnosis using robust inference system. International Journal of Medical Informatics, 2012.
- [5] Shahbakhhi M, Far D, Tahami E. Speech analysis for diagnosis of Parkinson's disease using genetic algorithm and Support Vector Machine, 2014.
- [6] Das, Resul. A comparison of multiple classification methods for diagnosis of Parkinson disease. Expert Systems with Applications, 37.2 2010, p. 1568-1572.

- [7] Little, Max A. et al. Suitability of dysphonia measurements for telemonitoring of Parkinson's disease. Biomedical Engineering, IEEE Transactions on 56.4 2009, p. 1015-1022.
- [8] Sakar B, Kursun O. Collection and analysis of a Parkinson speech dataset with multiple types of sound recordings, 2013.
- [9] Gök M. An ensemble of k-nearest neighbours algorithm for detection of Parkinson's disease. International Journal of Systems Science, 2013.
- [10] Kavzoğlu, Taşkın, Çölkesen İ. Destek vektör makineleri ile uydu görüntülerinin sınıflandırılmasında kernel fonksiyonlarının etkilerinin incelenmesi. Harita Dergisi 144.7 2010, p. 73-82.
- [11] Cortes, Corinna, and Vladimir Vapnik. Support-vector networks. Machine learning 20.3 1995, p. 273-297.
- [12] Theodoridis S, Koutroumbas K. Pattern Recognition, Fourth Edition: Academic Press, 2008.
- [13] Bahrepour, Majid, et al. Sensor fusion- based activity recognition for Parkinson patients. 2011, p. 171-190.
- [14] Solmaz R, Günay M , Alkan A. Fonksiyonel Tiroit Hastalığı Tanısında Naive Bayes Sınıflandırıcının Kullanılması , KSÜ, Kahramanmaraş, Türkiye.
- [15] Claire Nellec, Cine. Machine Learning: ECML-98, 10th European Conference on Machine Learning Chemnitz, Germany, April 21-23, 1998.
- [16] Cingiz M, Albayrak A, Amasyalı M. Evaluation of robustness of ensemble learners to noisy data. Bilgisayar Mühendisliği Bölümü Yıldız Teknik Üniversitesi, İstanbul, Türkiye.
- [17] Rodriguez J, Kuncheva I, Alonso C. Rotation Forest: A New Classifier Ensemble Method. IEEE Transactions on Pattern Analysis and Machine Intelligence, v.28n.10, October 2006, p.1619-1630.