ZOO VERİ SETİ ÜZERİNDE kNN ALGORİTMASININ ÖĞRENME BAŞARISININ TESPİT EDİLMESİ

DETERMINING LEARNING SUCCESS of kNN ALGORITHM on ZOO DATASET

Ahmet ÇELİK

Kütahya Dumlupınar Üniversitesi, Tavşanlı Meslek Yüksekokulu, Bilgisayar Teknolojileri Bölümü, 43300, Kütahya, Türkiye

Geliş Tarihi / Received: 06.10.2021
Kabul Tarihi / Accepted: 22.11.2021

Araştırma Makalesi/Research Article
DOI: 10.38065/euroasiaorg.762

ÖZET

İnsanlar çevresini inceleyerek, gözlem yaparak, araştırarak öğrenmektedir. Bu öğrendiklerinden aslında tecrübe kazanmaktadır. Kazandıkları tecrübelerini kullanarak, karşılaştığı yeni duruma uyum sağlamakta ve kararlar verebilmektedir İnsanlar nesneleri tanımlarken, sınıflandırma yaparken hep önceki bilgileriyle kıyaslama yaparak kararlar vermektedir. Önceden öğrendiği nesnelere benzerlik ve farklılıklar karar vermede çok etkilidir. Tecrübeye dayalı öğrenme yönteminin makineler üzerinde de kullanılabileceği yapılan çalışmalarda gösterilmiştir. Yapısında makine öğrenme yöntemlerini kullanan akıllı makineler veya cihazlar birçok alanda yaygın olarak kullanılmaktadır. Makine öğrenmesi farklı algoritmalar kullanılarak gerçekleştirilebilmektedir. Bu algoritmalar karar verirken veri seti içindeki nesnelerin öz niteliklerini kullanmaktadır. Nesnelerin öz niteliklerindeki benzerlik ve farklılıklar önceki tecrübelerle kıyaslanarak elde edilmektedir. Kıyaslama sonucu, karar verilerek nesnelerin sınıfları hakkında tahminler oluşturulmaktadır.

Bu çalışmada, Zoo veri seti üzerinde denetimli öğrenme yöntemi olan kNN makine öğrenme algoritması kullanılmıştır. Bu veri setinde yaygın karşılaşılan canlıların öznitelikleri bulunmaktadır. Bu öznitelikler kullanılarak veri setindeki canlıların sınıfları belirlenmektedir. kNN algoritmasında seçilen “k” komşu değeri ve ağırlık parametresi öğrenme başarısını etkilemektedir. Yapılan bu çalışmada, kNN algoritmasında kullanılan iki parametrenin öğrenme başarısına etkisi gösterilmiştir. Elde edilen sonuçlara göre “k=1” komşu değeri ve “Distance Ağırlık” parametresi seçilerek en yüksek başarı sonucu elde edilmiştir.

Anahtar Kelimeler: Makine Öğrenmesi, kNN Algoritması, Sınıflandırma, Tahmin, Zoo Veri Seti, Ağırlık Parametresi.

ABSTRACT

People learn by examining, observing and researching their environment. They actually gains experience from what they have learned. By using the experience they have gained, they can adapt to the new situation they encounter and make decisions. People always make decisions by comparing their previous knowledge while describing objects and classifying them. Similarities and differences to previously learned objects are very effective in decision making. It has been shown in the studies that the experiential learning method can also be used on machines. Intelligent machines and devices that use machine learning methods in their structure are widely used in many areas. Machine learning can be performed using different algorithms. These algorithms use the attributes of the objects in the data set when making decisions. Similarities and differences in the attributes of objects are obtained by comparing them with previous experiences. As a result of the comparison, a decision is made and predictions are made about the classes of the objects. In this study, kNN machine learning algorithm, which is a supervised learning method, was used on the Zoo dataset. In this data set, there are attributes of common living things. By using these attributes, the classes of living things in the data set are determined. The “k” neighbor value and weight parameter selected in the kNN algorithm affect the learning success. In this study, the effect of two parameters used in the kNN algorithm on learning success is shown. According to the results obtained, the "k=1"

neighbor value and the "Distance Weight" parameter were selected and the highest success result was obtained.

Keywords: Machine Learning, kNN Algorithm, Classification, Prediction, Zoo Dataset, Weight Parameter.

1. GİRİŞ

Makine öğrenme yöntemleri denetimli ve denetimsiz olarak incelenmektedir. Denetimli öğrenmede öğrenilmesi istenen sonuçlar bilinmektedir. Denetimsiz öğrenme de ise sonuçlar önceden bilinmemekte algoritmanın eğitim verilerini kullanarak sonuçları belirli sınıflara ayırması beklenmektedir. Denetimli öğrenme yöntemine pekiştirmeli öğrenme yöntemi de denilmektedir. Denetimli öğrenme algoritmaları olarak k en yakın komşu (KNN), Naive Bayes (NB), Karar Ağacı (DT) ve Destek Vektör Makinesi (SVM) yaygın olarak kullanılmaktadır.

Kanju vd. (2019), kullanıcıların bilgilerini çalan, güvensiz form bileşenlerine sahip olan web sitelerinin tespit etmek için kNN Algorithm, Naive Bayes, Decision Tree, Support Vector Machines, Neural Network and Random Forest makine öğrenme yöntemlerini kullanmışlardır.

Çelik (2020), COVID-19 Survilance veri setindeki hasta semptom bilgilerini kullanarak, COVID-19 hastalık teşhisini, Apriori birliktelik kuralı algoritmasını yardımıyla tahmin edebilecek bir çalışma gerçekleştirmiştir.

Zafar (2020), Arab el yazma eserlerini sınıflandırmak için kNN ve SVM algoritmalarını kullanmışlardır. Sınıflandırması yapılacak görüntülerin özelliklerini çıkarmak için Histogram of Oriented Gradients (HOG) and Local Binary Pattern (LBP) öz nitelik çıkarma yöntemlerini kullanmışlardır.

Yigit (2013), k en yakın komşu (kNN) algoritmasında kullanılabilecek yeni bir ağırlık parametresi geliştirmiştir. kNN algoritmasında kullanılan ağırlık parametresi de başarı performansını etkilemektedir.

kNN az hesaplama işlemi gerektiren, basit çalışma yöntemi olan denetimli bir makine öğrenme algoritmasıdır. Bu algoritmada yeni girilen nesne, veri seti içinde yakın olduğu k adet komşu değerden oylama sonuçlarına göre en yüksek oy oranına sahip olan sınıfa dahil edilmektedir (Goel 2017).

Bu çalışmada, Zoo veri seti içinden 50 adet test verisi ve 51 adet eğitim verisi seçilerek kNN makine öğrenme algoritmasının performansı ölçülmüştür. Algoritma da Öklid uzaklık vektörü kullanılarak, k komşu değerlerinin değişimi ve ağırlık parametre değerlerinin “Uniform” ve “Distance” seçilerek performansa etkisi test edilmiştir.

2. MATERYAL ve YÖNTEM

Bu çalışmada pekiştirmeli öğrenme yöntemlerinden biri olan kNN(k En Yakın Komşu) algoritması kullanılmıştır.

2.1. kNN komşu algoritması

K-Nearest Neighbor (KNN) algoritması, 1968 yılında Cover ve Hart isimli araştırmacılar tarafından geliştirilmiştir (Du et al. 2018). kNN algoritması, denetimli bir makine öğrenme yöntemi olduğundan önceden sınıfları belirlenmiş veri setleri üzerinde uygulanmalıdır. Yeni nesnenin öz nitelikleri kullanılarak sınıflara olan mesafesi Öklid vektörü kullanılarak hesaplanmaktadır. k tane komşunun çoğunluk oylaması yapılarak yeni nesne sınıfta olacağı tahmin edilmektedir Silahtaroğlu 2016). Denklem 1 üzerinde Öklid uzaklık vektör hesaplaması gösterilmektedir.

$$d_{Euclidean}(A_i, B_i) = \sqrt{\sum_{i=1}^{n}(A_i - B_i)^2} \quad 1)$$

A_i sınıfı tahmin edilecek i'inci örneği göstermektedir. B_i ise veri seti içindeki sınıfı belli olan örneklerdir. n öznitelik sayısıdır. $d_{Euclidean}(A_i, B_i)$, ise A_i ve B_i arasındaki uzaklık değeridir. Bu denklem kullanılarak k tane komşu bulunur. Komşu verilerin çoğunluğu hangi sınıfa aitse A_i örneği o sınıf olacağı tahmin edilir (Du et al. 2019, Balaban and Kartal 2018).

Öklid vektöründeki "Uniform" ağırlık seçildiğinde yeni veri, tüm noktalarıyla eşit ağırlık değeriyle karşılaştırılmaktadır. Eğer "Distance" ağırlık parametresi seçildiğinde, yeni verinin yakın komşu değerleri, uzaktaki komşulardan daha büyük etkiye sahip olmaktadır.

2.2. Zoo Veri seti

Bu veri seti Richard S. Forsyth tarafından 1990 yılında oluşturulmuş 16 özelliğe göre hayvanların 7 sınıfa ayrılması gerçekleştirilmiştir. Bu veri setinde 101 hayvana ait bilgiler vardır. Tablo 1 üzerinde veri setinin sınıfları verilmektedir. Bu veri setine UCI depo üzerinden ulaşılabilmektedir. Bu çalışmada, verilen eğitim seti içindeki veriler kullanılarak test verilerinin sınıflarının tahmin başarıları ölçülmüştür.

Tablo 1. Zoo veri setindeki sınıflar

Sıra	Sınıf Adı
1	Omurgasız
2	Böcek
3	Yüzergezer
4	Sürüngen
5	Memeli
6	Kuş
7	Balık

Bu sınıfları tespit edebilmek için Tablo 2 üzerinde Zoo veri setinin karakteristik özellikler(öz nitelikler) kullanılmaktadır. Bu özelliklerin 15 tanesinin veri içeriği Var/Yok biçimindedir. Ver türünün "Var" içeriği "1" ve "Yok" içeriği "0" olarak işlem alınmaktadır. Ayrıca "Ayak Sayısı" öz niteliğinin içeriği ise 0-6 arasında değişen sayısal verilerden oluşmaktadır. "Tip" alanı ise Tablo 1 verilen sınıf değerlerini göstermektedir.

Tablo 2. Zoo veri seti karakteristikleri

Sıra	Hayvan Özellikleri	Veri Tipi
1	Saç	Var/Yok
2	Tüy	Var/Yok
3	Yumurta	Var/Yok
4	Süt	Var/Yok
5	Uçar	Var/Yok
6	Yüzer	Var/Yok
7	Yırtıcı	Var/Yok
8	Dişli	Var/Yok
9	Omurga	Var/Yok
10	Nefes Alır	Var/Yok
11	Zehirli	Var/Yok
12	Yüzgeç	Var/Yok
13	Ayak Sayısı	0-2-4-6

14	Kuyruk	Var/Yok
15	Yerli	Var/Yok
16	Kedi Boyutu	Var/Yok
17	Tipi	Sınıf

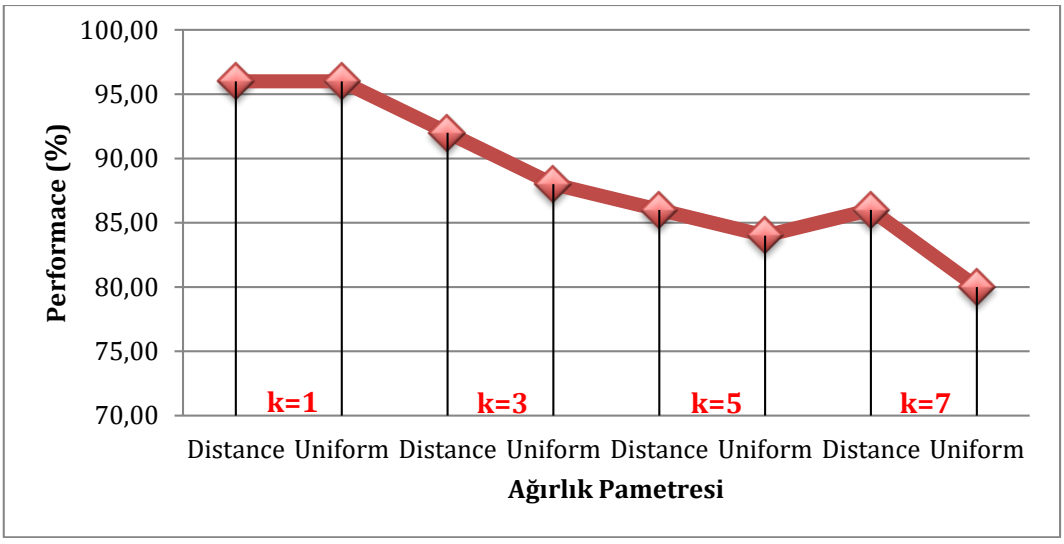
3. BULGULAR ve TARTIŞMA

Yapılan çalışmada veri seti içinden 50 adet test verisi rastgele seçilmiştir. Eğitim amaçlı ise 51 adet veri kullanılmıştır. Eğitim ve test verilerinin sınıflara eşit dağılımda olmasına dikkat edilmiştir. Elde edilen test sonuçları Tablo 3 üzerinde gösterilmektedir. Performansın k komşu değerine ve Öklid ağırlık parametresine göre değiştiği görülmüştür.

Tablo 3. kNN algoritmasının k komşu değeri ve ağırlık parametre seçimine bağlı olarak performans değerleri

k Değeri	Ağırlık Parametresi	Doğru Tahmin	Yanlış Tahmin	Performans
1	Distance	48	2	96,00
1	Uniform	48	2	96,00
3	Distance	46	4	92,00
3	Uniform	44	6	88,00
5	Distance	43	7	86,00
5	Uniform	42	8	84,00
7	Distance	43	7	86,00
7	Uniform	40	10	80,00

Tablo 3 den elde edilen sonuçlara göre şekil 1 deki grafik elde edilmiştir. En yüksek başarı performansının %96 değeriyle k komşu değeri 1 seçildiğinde elde edilmiştir. k komşu değeri arttıkça performansın düştüğü görülmektedir. En düşük başarı performansı %80 değeriyle k=7 seçildiğinde elde edilmiştir. Ayrıca Ağırlık parametresinin de performansa etkisi görülmüştür. Ağırlık parametresi "Distance" seçildiğinde, k komşu değerinin 1 haricindeki değerlerinde "Uniform" parametresine göre daha iyi sonuç verdiği görülmektedir.



Şekil 1. k komşu değeri ve ağırlık parametre seçime göre başarı performansının değişim grafiği

4. SONUÇLAR

Bu çalışmada, kNN denetimli makine öğrenme algoritmasının Zoo veri seti üzerinde değişen, k komşu değerine ve Öklid ağırlık parametresine göre başarı performansı ölçülmüştür. Elde edilen sonuçların uygun k komşu değerinin seçilmesinin çok önemli olduğunu ayrıca Öklid ağırlık parametresinde performansa etkisinin olduğunu göstermiştir.

Yapılan çalışmada, Zoo veri setindeki 101 veriden eğitim seti olarak, sınıflara eşit dağılım yapacak şekilde 51 tane ve test verisi olarak 50 tane veri seçilmiştir. Bu 50 test verisinin k komşu değerinin 1, 3, 5 ve 7 olduğu ve aynı zamanda ağırlık parametresinin "Distance" ve "Uniform" olduğu durumlardaki test başarısı test edilmiştir. Uygulama aracı olarak Orange kullanılmıştır. Bu araç birçok makine öğrenmesi ve veri madenciliği aracını yapısında barındırmaktadır.

Elde edilen sonuçlar, Zoo veri seti için en uygun k komşu değerinin 1 olduğunu göstermiş ve k komşu değerinin arttığında tahmin başarısının düştüğünü göstermiştir. Ayrıca Öklid "Distance" ağırlık parametresinin, "Uniform" ağırlık parametresinde göre daha yüksek bir başarı sağladığı görülmüştür. Bu çalışma, kNN makine öğrenmesinin tahmin başarını etkileyen faktörleri açık şekilde göstermektedir. Bundan sonraki çalışmalara temel teşkil edebilecektir.

KAYNAKLAR

Goel, A. and Mahajan S (2017). Comparison: KNN & SVM Algorithm. International Journal for Research in Applied Science & Engineering Technology (IJRASET), 5(13), 165-168.

Balaban, M. E. ve Kartal, E (2018). Veri madenciliği ve makine öğrenmesi temel algoritmaları ve R Dili ile Uygulamalar (2. Basım), İstanbul, Türkiye: Çağlayan Kitap & Yayıncılık & Eğitim, 48-72.

Du, S. and Li, J (2019). Parallel Processing of Improved KNN Text Classification Algorithm Based on Hadoop. 2019 7th International Conference on Information, Communication and Networks (ICICN) (pp. 167-170), doi: 10.1109/ICICN.2019.8834973.

Kunju, M. V., Dainel, E., Anthony H. C. and Bhelwa, S (2019). Evaluation of Phishing Techniques Based on Machine Learning. 2019 International Conference on Intelligent Computing and Control Systems (ICCS) (pp. 963-968), doi: 10.1109/ICCS45141.2019.9065639.

Silahtaroğlu, G (2016). Veri madenciliği (Kavram ve algoritmaları) (3. Basım). İstanbul, Türkiye: Papatya Yayıncılık Eğitim, 118-120.

Yigit, H (2013). A weighting approach for KNN classifier. 2013 International Conference on Electronics, Computer and Computation (ICECCO) (pp. 228-231).

Demsar J., Curk T., Erjavec A., Gorup C., Hocevar T., Milutinovic M., Mozina M., Polajnar M., Toplak M., Staric A., Stajdohar M., Umek L., Zagar L., Zbontar J., Zitnik M., Zupan B (2013). Orange: Data Mining Toolbox in Python, Journal of Machine Learning Research 14(Aug), 2349−2353.

Dua, D. and Graff, C., UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]". Irvine, CA: University of California, School of Information and Computer Science; 2019.

Zafar A and Iqbal, A (2020). Machine Reading of Arabic Manuscripts using KNN and SVM Classifiers. 2020 7th International Conference on Computing for Sustainable Global Development (INDIACom) (pp. 83-87), doi: 10.23919/INDIACom49435.2020.9083696.